REFERENCE ONLY

UNIVERSITY OF LONDON THESIS

Degree PhD Year 2006 Name of Author BREWER D.S

COPYRIGHT

This is a thesis accepted for a Higher Degree of the University of London. It is an unpublished typescript and the copyright is held by the author. All persons consulting the thesis must read and abide by the Copyright Declaration below.

COPYRIGHT DECLARATION

I recognise that the copyright of the above-described thesis rests with the author and that no quotation from it or information derived from it may be published without the prior written consent of the author.

LOANS

Theses may not be lent to individuals, but the Senate House Library may lend a copy to approved libraries within the United Kingdom, for consultation solely on the premises of those libraries. Application should be made to: Inter-Library Loans, Senate House Library, Senate House, Malet Street, London WC1E 7HU.

REPRODUCTION

University of London theses may not be reproduced without explicit written permission from the Senate House Library. Enquiries should be addressed to the Theses Section of the Library. Regulations concerning reproduction vary according to the date of acceptance of the thesis and are listed below as guidelines.

A. Before 1962. Permission granted only upon the prior written consent of the author. (The Senate House Library will provide addresses where possible).

B. 1962 - 1974. In many cases the author has agreed to permit copying upon completion of a Copyright Declaration.

C. 1975 - 1988. Most theses may be copied upon completion of a Copyright Declaration.

D. 1989 onwards. Most theses may be copied.

This thesis comes within category D.

☑ This copy has been deposited in the Library of UCL

☐ This copy has been deposited in the Senate House Library, Senate House, Malet Street, London WC1E 7HU.

Modelling the p53 Gene Regulatory Network

Daniel Simon Brewer

A thesis submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy
at the
University of London.

Centre for Mathematics and Physics in the Life Sciences and Experimental Biology
and
Institute of Child Health
University College London

2006

UMI Number: U592648

All rights reserved

INFORMATION TO ALL USERS
The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript
and there are missing pages, these will be noted. Also, if material had to be removed,
a note will indicate the deletion.

UMI U592648
Published by ProQuest LLC 2013. Copyright in the Dissertation held by the Author.
Microform Edition © ProQuest LLC.
All rights reserved. This work is protected against
unauthorized copying under Title 17, United States Code.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

For Alice and my parents

Acknowledgements

I would first and foremost like to thank Prof. Jaroslav Stark, Dr. Mike Hubank and Prof. Robin Callard for allowing me the opportunity to work with them. I am grateful for all their supervision, input into my work and improvements suggested for this manuscript. Special thanks go to Robin and Andy Yates for monthly meetings during the dark days that enabled me to get out of a rut and find a global perspective of my work. Also thanks to Mike who worked tirelessly to try and get me decent protein data in the latter stages. I am indebted to Martino Barenco for his supervision on the G time series and general willingness to discuss problems, to Andy for initial discussions on the localisation model and to David Bogle for advice on optimisation. This PhD was made possible by funds from the MRC and the ICH.

I would also like to thank the members of my year group in CoMPLEX: Olivier Cinquin, Dave Dale, Christian Bottomley, Simon Moon and Chris Mullaley. They are a great bunch and have provided extensive scientific interaction as well as social larks. I knew that we would have a great time from that very first night out! Special thanks go to Olivier for discussions on my localisation model and Apple related nonsense, Christian for his in depth advice on statistics, Chris for his C++ knowledge, and Simon & Dave for discussions on optimisation and other general moans. I would also like to thank the rest of CoMPLEX with whom I have interacted, for making my stay here enjoyable.

There are several people without whom I would not be writing this thesis at all. Most of all I am extremely grateful for Alice, who has been by my side through it all. She somehow managed to push me though even the bleakest of moments. I am glad I am on this journey with you mate ...it looks like we might have made it. My parents have been extremely important - with their relaxed attitude, love and support they have really helped. I would also like to acknowledge them for their sacrifice to get me a decent education. Also I am grateful for my siblings, Adam and Em - I could not dream of having a better pair.

Finally, I would like to thank all the others who are not directly connected with this work but have improved my life with their friendship (you know who you are). Particular mention goes to those that have provided me with great weekends away (Nic & Jill, and Dave & Alice) and the Imperial guard (Avnish, Jacob, Matt, Mike, Tessa, Paul and Jenny).

Abstract

p53 is the central protein in the DNA damage response and is part of a complex and extensive gene regulatory network. This network integrates a variety of stress signals to produce the up-regulation of active p53 and a range of effects including apoptosis, growth arrest and DNA damage repair. The p53 system has typically been studied qualitatively as a linear pathway, however this approach is insufficient to gain a full functional understanding of the dynamic nature of the network. In this work a better description of the DNA damage response will be constructed through the use of mathematical techniques.

Ordinary differential equations models of the p53 network between DNA damage and p53 up-regulation are proposed, including a model that takes into account various localisation mechanisms. Parameter estimation is required to validate these models with biological data. A number of established techniques are examined along with a novel method based on linear algebra, collocation and B-splines. To examine the network downstream of p53 and the global response to DNA damage, a "G" time profile ($G_g(t)$) quantifying the activity driving the formation of each gene is constructed. This is derived from a model of gene transcription, microarray data and mRNA degradation rates.

The new parameter estimation technique developed works significantly better than the other techniques examined. Also, it was found that the mechanisms that control the location of p53 significantly contribute to the rapid DNA damage response. The G time profiles suggest that there are four principal transcription activities in the DNA damage response: p53, an early peaking response (possibly AP-1), stopping and restarting the cell cycle, and a double peaked response. The G time profile in combination with a training set of genes can be used to successfully find confirmed p53 targets.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

When cells are stressed or damaged they can pose a threat to the organism *via* DNA damage. In the least threatening situation the cells fail to perform their function but still consume resources, and in the most serious cases the cells become cancerous. To safeguard against this threat there are systems that stop the cell from dividing until repairs can be made to the DNA, and other mechanisms that cause a tightly regulated cell suicide known as programmed cell death or apoptosis (Alberts *et al.*, 2002). Apoptosis is triggered if the cell has sustained so much damage that repair would be prone to too much error and therefore risk deleterious mutation. During apoptosis, normally dormant proteases called caspases are activated and cause the destruction of cellular proteins, resulting in cell death (Rich *et al.*, 1999). Caspases are activated either by an extracellular signal (*via* cell surface receptors known as death receptors), or through the release of cytochrome c from the mitochondria and the subsequent formation of a catalytic apoptosome (Alberts *et al.*, 2002). p53, the central protein in the DNA damage response can activate both pathways (Vogelstein *et al.*, 2000).

p53 is a tumour-suppressor and has been described as the "guardian of the genome" (Lane, 1992). It is part of a complex and extensive gene regulatory network[1] that integrates a variety of stress signals to produce a range of effects including apoptosis, growth arrest and DNA damage repair (see chapter 2). Of particular importance is p53's role in the decision to commence apoptosis, which is not well understood. p53 is known to play a vital role in preventing cancer; p53 is dysfunctional in the majority of cancer types (Soussi *et al.*, 2000) and more than 18000 different p53 mutations have been found in cancers (Bode and Dong, 2004). A greater understanding of the p53 gene regulatory system is likely in the long term to lead to an improvement in human health.

The p53 gene regulatory network is an extremely well studied system and numerous components and interactions have been discovered by using traditional techniques. Despite this, the examination of uni-directional pathways qualitatively is insufficient to gain a full functional understanding of the dynamic nature of the network. This is due to the complexity of the p53 gene regulatory network, which has numerous components, many feedback loops and interacts with a large number of cellular subsystems. As John Maddox (former Nature editor) states,

> "If some intermediate goal in biology is the understanding of the functioning of an entire cell, it is unthinkable that it will be attainable without quantitative information about the abundance of the component molecular species." (Maddox, 1992)

To gain functional insight into the DNA damage response a quantitative and predictive description of the network is needed. This requires mathematical modelling, quantitative data and advanced data analysis techniques.

[1] A gene regulatory network is a collection of genes that interact through both DNA segments and protein products to regulate the transcription rate of the component genes.

There are considerable challenges to modelling the p53 network, these range from the decision as to which components should be included in the model to best represent the system, to the choice of appropriate parameter values. Until very recently there have only been two models of the p53 gene regulatory network (Bar-Or *et al.*, 2000; Monk, 2003a). At the population level, experimental data suggests that there is damped oscillations between the two core components after DNA damage (Bar-Or *et al.*, 2000). Both models successfully replicated this dynamic but are limited in the number of components they examined and by the heuristic approach to choosing parameters (these are examined further in section 2.4.1 and 2.4.2). As the work on this thesis was near completion, two different models (Ciliberto *et al.*, 2005; Ma *et al.*, 2005) were published that replicate experiments that indicate that p53 pulses at the single cell level (Lahav *et al.*, 2004). These models were more extensive including more than the two core components.

In recent years there has been an explosion of quantitative mRNA data due to the development of microarrays which can simultaneously measure the expression levels of thousands of genes. Devising methods that can extract useful biological information from this vast amount of data is a major challenge. This is a general problem but also applies directly to the DNA damage response and in particular to detecting genes that are transcribed by p53. There has been no use of mathematical models to extract p53 targets from microarray data apart from a recent paper produced by this group (Barenco *et al.*, 2005).

In this work the overall aim is to gain a better description of the DNA damage response through the use of mathematical techniques. More specifically this work aims to achieve a better description of the p53 gene regulatory network between DNA damage and p53 up-regulation, identify targets of the p53 network and develop techniques that will help meet these aims. If possible the techniques developed should be generally applicable to other similar problems.

The experimental system used examines the DNA damage response in the p53 wild type human lymphoid cell line MOLT 4. Cells were exposed to ionising radiation and time series measurements were gathered for mRNA and various proteins (see chapter 3). The thesis will be divided into three parts.

In the first part of the thesis ordinary differential equations models of protein expression in the system are proposed based on previous biological knowledge. In the p53 system the regulation of the location of the core components appears to be important. Therefore, a model is proposed that takes into account some of these localisation mechanisms. This model is examined to determine whether the inclusion of localisation regulation improves the performance of the system.

The second part describes studies into parameter estimation using a small experimental protein data set. To make predictions based on these models it is important to determine how well the behaviour of the model matches biological data, i.e. to find the parameters that cause the model solutions to best fit the data. This is important as it

is rare to have direct parameter measurements in this kind of system. Established techniques such as simulated annealing are implemented and examined along with a novel technique based on linear algebra, collocation and a series of B-splines. A number of the proposed models are evaluated based on the latter technique.

It is easier to gather data on the amount of mRNA than the amount of functionally active protein. In the final part of the thesis, a simple model of positively regulated gene expression is investigated with a view to gaining information about the protein level response to DNA damage using microarray mRNA data. For each gene, a quantity called the G time profile is proposed which is representative of the time profile of the "activity" that is driving the quantity of that gene's mRNA. This quantity can be used to group genes that share the same transcription factor and find the main "activities" that drive the DNA damage response.

This thesis has a number of important results. It was found that the simple model of gene transcription does generate useful information about the protein level response from microarray data. In particular, the G time profile was used to find the principal transcription activity time profiles of the DNA damage response. When there is a high level of damage, it was found that there are four principal transcription activities: p53, AP-1, stopping and restarting the cell cycle, and a double peaked response (section 8.6.3). Also, the G time profile in combination with a training set of genes can be used to successfully find confirmed p53 targets (section 8.4).

Another key result is that the new parameter estimation technique developed worked extremely well when there was a small amount of error in the data even with small amounts of data and reasonably well when the amount of data was large and the error was large (chapter 7). Finally, it was found that the regulatory mechanisms in the p53 system that control the location of p53 significantly contribute to the rapid response that occurs after DNA damage (chapter 5).

Chapter 2

The p53 gene regulatory network

2.1 Introduction

The p53 gene is a tumour-suppressor gene (Vogelstein *et al.*, 2000) and has been described as "guardian of the genome" (Lane, 1992). It is known to play a vital role in the cell because when it is not functioning correctly cancer results; p53 is dysfunctional in the majority of cancer types (Soussi *et al.*, 2000) and greater than 18000 different p53 mutations have been found in cancers (Bode and Dong, 2004). Additionally, mice genetically engineered to lack the p53 gene show a shortened lifespan because of increased susceptibility to tumours (Hickman *et al.*, 2002).

The p53 protein is the major node of a network that works to apply the "brakes" on cell multiplication and in certain cases causes apoptosis. It's primary function is to act as a transcription factor. In a normal unstressed cell the concentration of p53 is kept very low. When stress occurs, the members of the network interact to produce a 3-10 fold increase in the concentration of activated p53 (Hickman *et al.*, 2002; Harris and Levine, 2005). This process is very quick with p53 levels increasing within minutes and the first apoptotic events occurring within a few hours in some cell types (Clarke *et al.*, 1994; Merritt *et al.*, 1994). There are three main routes that stimulate the increase in p53:

- DNA damage

  DNA damage can be lethal if left unrepaired. Furthermore mis-repair of damaged DNA can lead to the production of mutant proteins, which can contribute to cells becoming cancerous. Therefore it is important that the cell takes action when the DNA becomes damaged. There are two different DNA damage response pathways:

  1. ATM (ataxia telangiectasia mutated) is the key protein that signals double-stranded breaks in DNA (see section 2.2.2). This pathway is extremely sensitive and it has been suggested that a single double-stranded break in DNA may be sufficient to trigger a rise in levels of p53 (Vogelstein *et al.*, 2000). Double-stranded breaks in DNA are caused by, among other things, ionising radiation.
  2. DNA damage caused by a wide range of chemotherapeutic drugs, ultraviolet light and protein-kinase inhibitors triggers a pathway that depends on ATR (ataxia telangiectasia related) and casein kinase rather than ATM, CHK2 or $p14^{ARF}$ to stimulate the p53 network (Vogelstein *et al.*, 2000) ($p14^{ARF}$ is equivalent to the protein $p19^{ARF}$ in mice and from here on will be referred to as ARF).

- Aberrant growth signals

  This pathway is stimulated by the over expression of oncogenes such as RAS or Myc (Wahl and Carr, 2001). Oncogenes stimulate cell growth and so over-expression indicates that the cell has become cancerous necessitating cell reaction. In humans,

this over-expression of oncogenes is detected by ARF, which stimulates the p53 network (Sherr and Weber, 2000).

In this thesis the focus will be on the double-stranded break DNA damage pathway. For p53 to become functionally active it requires a conformation change so that it can form tetramers and bind strongly with DNA. This alteration is caused by the addition or removal of various chemical groups (Vogelstein *et al.*, 2000). In its active form the stability of p53 increases; the half-life of p53 increases from 6–20 mins to an hour (Harris and Levine, 2005).

Once the p53 protein is stabilised and activated it accumulates in the nucleus and binds to specific DNA sequences and promotes or suppresses the transcription of adjacent genes. p53 is very prolific targeting over 150 genes (Bode and Dong, 2004). There are several classes of protein that predominate in the p53 target profile and have tumour-suppressing effects (Vogelstein *et al.*, 2000):

- Apoptosis

  p53 transcribes a large number of proteins that are involved in both the intrinsic and extrinsic apoptotic pathways (Vousden, 2000). These include proteins from the Bcl2 family, death domain proteins, proteins that induce reactive oxygen species, the plasma membrane protein p53 apoptosis effector related to PMP-22 (PERP), the survival factor antagonist insulin-like growth factor binding protein 3 (IGF-BP3) and the apoptosis protease activator apoptotic activating factor one (APAF1) (Wahl and Carr, 2001). Bax, a member of the Bcl2 family, was the first apoptotic factor to be identified as a target for p53 transactivation (Hickman *et al.*, 2002).

- Growth arrest

  p53 transcribes genes that contribute to blocking the cell division cycle. It has been shown that p53 is essential for prolonged cell cycle arrest induced by ionising radiation, but that shorter, earlier cell cycle arrest is p53 independent (Wahl *et al.*, 1997). $p21^{Waf1/Cip1}$ is a protein transcribed by p53 that stands out as playing a critical role in the maintenance of cell cycle arrest (el Deiry *et al.*, 1993). It is believed that the short periods of cell stagnation are caused by CHK1 binding to Cyclin D1, with ATM possibly activating CHK1.

- Angiogenesis inhibition

  To sustain tumour growth there needs to be an influx of raw materials which requires contact with a large number of blood vessels. To aid this, many tumours send out chemical signals that encourage the growth of new blood vessels. It has been shown that in the presence of p53 the formation of new blood vessels is inhibited in tumours by the activation or repression of genes that regulate new blood vessel formation (Bouvet *et al.*, 1998; Dameron *et al.*, 1994)

- Genetic stabilisation

  p53 also participates in DNA damage repair, even though the mechanisms are not clearly understood. p53 transcribes target genes such as DDB2 (p48), GADD45 (Smith *et al.*, 1994) and p53R2 (Tanaka *et al.*, 2000), which are important in the regulation of nucleotide-excision repair of DNA, chromosomal recombination and segregation (Vogelstein *et al.*, 2000). It has been shown that cells lacking in p53 do not display nucleotide excision repair (Wani *et al.*, 1999) and base excision repair is less efficient (Offer *et al.*, 2001). It has also been suggested that p53 itself plays a role in genetic stability with the C-terminus of p53 binding to different forms of DNA damage (Balint and Vousden, 2001).

2.2 The major members of the network between damage and p53 up-regulation

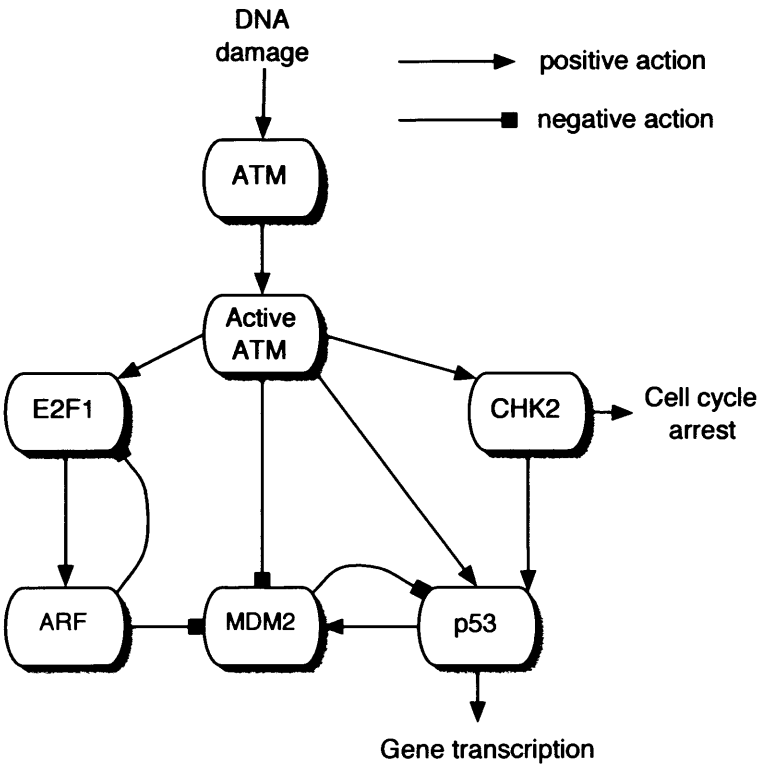


Figure 2.1: A diagram to summarise the major components and interactions in the protein activation side of the p53 DNA damage control network.

In this section the core and best known proteins that regulate the levels of p53 will be examined (Figure 2.1). These core components will be used as the basis for the models. The p53 DNA damage network is much more complex then these central features. Some of this further detail will be examined in section 2.3.

2.2.1 The core of the network

p53

The p53 gene was first described in 1979, when it was incorrectly identified as an oncogene and it was only ten years later that its true function as a tumour suppressor gene was discovered (Vogelstein *et al.*, 2000). It was first demonstrated by Yonish-Rouach *et al.* (1991) that DNA breaks caused a rise in p53 which in turn causes apoptosis.

The p53 protein consists of 393 amino acids and is commonly divided into three functional domains. the N-terminal, the central core and the C-terminal (Bode and Dong, 2004). The N-terminal consists of a domain required for transcriptional activity and another required for the interaction with the histone deacetylase, SIN3. The C-terminal contains a tetramerisation domain and two domains that control p53 localisation. It also contains a region at the very end of the C-terminal that can be used to regulate the core domain. Finally, the central core, which is the largest functional unit and contains the DNA binding region.

Although the ability of p53 to function as a transcription factor is considered to be its main function in apoptosis (Vogelstein *et al.*, 2000; Wahl and Carr, 2001; Harris and Levine, 2005), recent evidence has suggested p53 has other transcriptionally independent apoptotic functions (Vousden, 2005; Schuler and Green, 2005). It has been found that p53 acts functionally as a BH3-only protein (Erster and Moll, 2005), a family of proteins that enable the activation of pro-apoptotic proteins Bax and/or Bak. BH3-only proteins come in two classes: activators that directly bind and activate Bax and Bak, and enablers which form a complex to anti-apoptotic proteins such as BCL2 or BCL-XL, freeing activators to perform their function (Yee and Vousden, 2005). It has been shown that p53 is both an activator (Chipuk *et al.*, 2004, 2005; Leu *et al.*, 2004) and an enabler in the mitochondria (Mihara *et al.*, 2003). There is still debate about the importance of these transcriptionally independent activities of p53; some experiments have show that apoptosis can still occur in certain cell types when there is a p53 mutant that cannot activate transcription (Hickman *et al.*, 2002) whereas other experiments demonstrate loss of cell cycle arrest and apoptotic functions (Jimenez *et al.*, 2000; Chao *et al.*, 2000).

MDM2

Mouse double minute two (MDM2 or sometimes known as HDM2 in humans) is an ubiquitin ligase that is the key negative regulator of p53 (Wahl and Carr, 2001). The importance of MDM2 in the control of p53 has been well established. In mice without MDM2 present there is unrestrained p53 activity which blocks normal growth and development - this causes early embryonic lethality (Jones *et al.*, 1995). The survival of mice lacking in p53 as well as MDM2 demonstrate the co-dependent nature of the system (Jones *et al.*, 1995). When MDM2 is over-expressed there is an associated development of tumours that do not mutate wild type p53 (Oliner *et al.*, 1992), which suggests that

MDM2 can interfere with the sensitivity of the p53 network to cell stress.

MDM2 regulates p53 through ubiquitin-mediated proteolysis (Haupt *et al.*, 2003). MDM2 attaches several ubiquitin molecules to the p53 protein and these act as a label to the protesome machinery to degrade the p53 protein (Vogelstein *et al.*, 2000). Efficient degradation of the p53 protein requires export from the nucleus to the cytoplasm where the majority of protesomes are located (although some degradation still occurs within the nucleus) (Balint and Vousden, 2001). Ubiquitination contributes to the efficient nuclear export of p53 (Boyd *et al.*, 2000; Geyer *et al.*, 2000) possibly by changing the form of the p53 protein to allow access to the nuclear export sequence (Stommel *et al.*, 1999). Removal of p53 protein from the nucleus represses its transcription activity. In certain tumours wildtype p53 is found trapped outside the nucleus suggesting the up-regulation of nuclear export possibly by an MDM2 mutant (Balint and Vousden, 2001).

MDM2 also suppresses p53 activity by binding to its N-terminal, directly blocking p53's ability to bind to DNA (Wahl and Carr, 2001; Chen *et al.*, 1996). MDM2 also inhibits acetylation of p53. Cofactors such as p300 bind to p53 and cause the acetylation of the C-terminal which negatively affects the ability of MDM2 to ubiquitinate p53 (Wahl and Carr, 2001). Hence by preventing acetylation, MDM2 ensures that p53 is less likely to stabilise and become active.

MDM2 itself is a p53 target gene (Barak *et al.*, 1993) and so p53 and MDM2 form a negative feedback loop. A rise in p53, will cause a rise in MDM2 which in turn will suppress p53 and reduce its amount. This negative feedback loop works to keep the equilibrium level of p53 low. When there is a stress placed on the cell these interactions are modified so that p53 levels can increase and activate the necessary pathways. This occurs in two main ways: the modification of the p53 protein or the modification of the MDM2 protein. Both of these negatively affect the binding of MDM2 to p53 and hence prevent the ubiquitination and degradation of p53. In p53 there are numerous phosphorylation sites within or near the N-terminal MDM2 binding region of p53 and phosphorylation at many of these sites can help prevent the binding of p53 to MDM2 (Balint and Vousden, 2001).

2.2.2 Main regulators of the core

ATM

Ataxia telangiectasia mutated (ATM) is a kinase that is the key signal transducer of the response to double-stranded DNA breaks (Abraham and Tibbetts, 2005). ATM is the gene defective in ataxia telangiectasia (a progressive neurodegenerative disease) patients who have a predisposition to cancer and extreme cellular radio-sensitivity (Savitsky *et al.*, 1995). In cells not under stress, ATM resides as an inactive dimer or higher order multimer (Bakkenist and Kastan, 2003). When double-stranded breaks occur, ATM is recruited to the DNA damage site (Smith *et al.*, 1999; Lavin *et al.*, 2005) and dissociates

becoming functionally active (Bakkenist and Kastan, 2003). The Mre11-Rad50-Nbs1 complex both recruits and activates ATM at the DNA damage sites (Lee and Paull, 2005; Falck *et al.*, 2005).

Having become functionally active, ATM transmits the damage signal. Among other proteins in the p53 damage network, active ATM phosphorylates MDM2 and p53 disrupting the negative feedback loop, allowing p53 levels to rise (Norbury and Zhivotovsky, 2004). ATM phosphorylates human MDM2 on Ser 395 (Maya *et al.*, 2001) which compromises MDM2's ability to inhibit and destabilise p53 (Goldberg *et al.*, 2002). ATM also phosphorylates p53 at Ser 15 (Canman *et al.*, 1998), which prevents MDM2 from binding to p53 but allows cofactors to bind encouraging the activation of p53. After UV damage an ATM related molecule (ATR) signals the damage.

CHK2

CHK2 is an important protein in the check point control system. It is activated through phosphorylation by ATM (Norbury and Zhivotovsky, 2004). Activated CHK2 phosphorylates Ser 20 of p53 which is within the MDM2 binding region (Chehab *et al.*, 1999, 2000). This prevents MDM2 binding and so has a positive effect on the amount of p53 present. CHK2 is important to the damage control system; mice lacking in CHK2 have a faulty p53-dependent cell cycle and apoptotic responses (Bell *et al.*, 1999). CHK2 also directly causes short term arrest in the cell cycle by phosphorylating the cell cycle regulator CDC25A (Falck *et al.*, 2001).

E2F1

E2F1 is a transcription factor (Rich *et al.*, 2000) that is phosphorylated and stabilised by CHK2 in response to DNA damage (Stevens *et al.*, 2003; Wahl and Carr, 2001). The stabilisation of E2F1 was found to be required for p53 to induce apoptosis in thymocytes (Lin *et al.*, 2001). E2F1 can both promote and inhibit cell growth. E2F1 is involved in the p53 network as a factor that promotes the transcription of ARF (Zhu *et al.*, 1999). There is some evidence that p53 inactivates E2F1 but the evidence is unconvincing (Sherr and Weber, 2000).

ARF

The alternative reading frame product (ARF) is a small protein that is one of two products of the Ink4a-Arf locus (Lowe and Sherr, 2003). It plays an important role in the DNA damage response; ARF-null mice exhibit almost the same tumour predisposition as p53-null mice (Kamijo *et al.*, 1997) and DNA damage increases the levels of ARF (Khan *et al.*, 2000). ARF is the centre of a complex network and has p53-independent anti-proliferative activities and is also mutated in many cancers (Lowe and Sherr, 2003). It plays a central role in the suppression of activated oncogenes.

ARF also directly regulates the core of the p53 network, binding directly to MDM2, which blocks both the ubiquitination of p53 by MDM2 and the inhibition of p53 acetylation by MDM2 (Balint and Vousden, 2001). This allows the stabilisation of p53. ARF also down-regulates E2F1 producing a negative feedback loop between ARF and E2F1 (Mason *et al.*, 2002). ARF, like p53, is expressed at very low levels in cells that are not under stress (Vousden, 2000).

2.3 The complexity of the p53 network

In the previous section the core components of the p53 network were introduced. This did not reveal the true complexity of the network as it is currently known. p53 itself is complex on many levels (genomic structure, regulation and function) (Braithwaite *et al.*, 2005) and the network of interactions are equally complex (Kohn and Pommier, 2005). In this section the main themes of this complexity will be briefly examined.

The p53 network behaves differently in different tissues and cell lines (Bouvard *et al.*, 2000; Fridman and Lowe, 2003). For example, splenic lymphocytes readily initiate apoptosis after exposure to ionising radiation but for cardiac myocytes apoptosis forms no part of the DNA damage response (Rich *et al.*, 2000). The importance of p53 even varies in different cell lines, a p53 mutant that cannot activate or repress transcription has been shown to be both *functional* and *deficient* in the induction of apoptosis, depending on the cell-culture system used (Hickman *et al.*, 2002).

One possible cause for these differences in output from the p53 network is that due to tissue specific expression, there are different components present in different tissues. Another possible arbitrator of the network output is the diverse array of covalent post-translational modifications that occur to p53 and markedly effect the expression of p53 targets (Bode and Dong, 2004). Phosphorylation and acetylation seem to be the most important p53 modifications. Phosphorylation of p53 generally stabilises it and increases its sequence specific binding (Hupp and Lane, 1994). p53 has at least 17 phosphorylation sites with significant redundancies, with different sites phosphorylated by multiple kinases and multiple sites phosphorylated by a single kinase. It has been speculated that the pattern of phosphorylation could determine the response of the cell to DNA damage (Bode and Dong, 2004). Acetylation is also important as it it stabilises p53 (Ito *et al.*, 2001) and affects the p53-dependent apoptotic response (Luo *et al.*, 2000). There are numerous cofactors involved in acetylation but p300, CBP and PCAF seem to have a significant effect on the performance of the system (Balint and Vousden, 2001).

Another complication is that both p53 and MDM2 are part of a larger family of proteins (Michael and Oren, 2002). The MDM2 family consists of one extra protein, MDMX. MDMX appears to play an important role in cancer; it is overexpressed in numerous tumours where there are elevated levels of wildtype p53 (Ramos *et al.*, 2001). It was also found that MDMX-deficient mice die as embryos, caused by p53-mediated

cell growth arrest (Parant *et al.*, 2001). MDMX does not target p53 for degradation but stabilises both p53 and MDM2 (Stad *et al.*, 2000). MDMX may act competitively with MDM2 preventing MDM2 binding with p53 or itself, thus preventing ubiquitination and hence degradation. p53 is a member of a family that has two additional proteins p63 and p73 (both are believed to be evolutionary precursors of p53). They share over 60% of amino acids in the core domain (Slee *et al.*, 2004) and have a large number of common genes that they regulate (Harms *et al.*, 2004). All three can induce apoptosis but p63 and p73 do not have other tumour suppressing effects. There are functional differences, on studies of mice deficient in p53, p63 and p73 it has been established that p63 and p73 are much more important than p53 for mouse development but that lack of p53 causes more tumour growth than the lack of p63 or p73 (Donehower *et al.*, 1992; Yang *et al.*, 1999, 2000). In addition mutations of p63 and p73 in tumours are rare which questions the importance of these two proteins in protecting the cell (Slee *et al.*, 2004). p63 and p73 will interact with p53, but it is still unclear whether p63 and p73 play a competitive, regulatory or redundancy role (Fridman and Lowe, 2003).

The p53 network is not isolated in a single cell but is affected by other cells through the AKT survival pathway (Haupt *et al.*, 2003). Growth and survival factors cause the activation of AKT through the function of PI3K (Lawlor and Alessi, 2001). AKT phosphorylates MDM2 which reduces the affinity of MDM2 to ARF, enhances the nuclear accumulation of MDM2 and encourages the interaction of MDM2 with p300 (Testa and Bellacosa, 2001). All these increase the interactions of MDM2 and p53 and hence increase the inhibition and degradation of p53. Once the cell senses stress, p53 tries to counteract this process through a number of mechanisms; p53 promotes the degradation of AKT (Gottlieb *et al.*, 2002), transcribes cyclinG which dephosphorylates MDM2 on the AKT phosphorylation sites (Okamoto *et al.*, 2002; Haupt *et al.*, 2003) and transcribes PTEN which suppresses the activity of PI3K (Mayo and Donner, 2002).

Recently, it has become clear that a key mechanism in the regulation of the p53 network response to stress is the control of the location of the network's principal components, in particular p53 and MDM2 (O'Brate and Giannakakou, 2003; Liang and Clarke, 2001; Michael and Oren, 2003). This issue will be examined in chapter 5.

Every year new potentially important proteins are discovered to regulate and interact with p53. Recently two other ubiquitin ligases have been discovered that target p53 for degradation and are transcriptionally activated by p53: COP1 (Dornan *et al.*, 2004) and Pirh2 (Leng *et al.*, 2003). Thus MDM2, COP1 and Pirh2 all have a similar negative feedback with p53. It is currently unclear why there is this redundancy (Harris and Levine, 2005). There are also proteins that deubiquitinate p53 such as HAUSP (Li *et al.*, 2002; Lim *et al.*, 2004), which has also been found to regulate MDM2 independently of p53 (Li *et al.*, 2004b). Another set of recently discovered regulators of p53 are the ASPP family of proteins (Slee *et al.*, 2004). These three proteins, ASPP1, ASPP2 and iASPP, have been shown to be important in determining whether pro-apoptotic genes

are transcribed by p53. ASPP1 and ASPP2 have been shown to increase the levels of pro-apoptotic p53 targets such as Bax and PIG3, but have minimal effect on the genes that affect other functions, such as MDM2, cyclinG and p21 (Samuels-Lev *et al.*, 2001). iASPP on the other hand behaves as an inhibitor of the rest of the family, and hence acts as an inhibitor of p53-dependent apoptosis (Bergamaschi *et al.*, 2003). It has been speculated that the relative levels of these proteins may determine what level of DNA damage is required for apoptosis to occur (Norbury and Zhivotovsky, 2004).

At this stage, it is impossible to model all the complexity that is present in the p53 network, therefore only the principal components that are known to be active in the biological system under study will be considered.

2.4 Models in the literature

There have only been four mathematical models of the p53 gene regulatory network based at the molecular level[1]. In this section these models will be examined along with some experimental results that suggest interesting dynamics.

2.4.1 A model of p53-MDM2 interaction that gives oscillations

Bar-Or *et al.* (2000) proposed a model of the p53 network that is fairly simple as it only has a few components but complicated as it takes into account many different relationships (Figure 2.2). The three "substances" modelled are p53, MDM2 and I, where I is a hypothetical intermediary state of p53 used to introduce a delay between p53 activation and p53 induction of MDM2. The rate of change in the concentration of the components are given by,

$$\begin{aligned}
\frac{dp53(t)}{dt} &= source_{p53} - p53(t) \times MDM2(t) \times degradation(t) - d_{p53}p53(t), \\
\frac{dMDM2(t)}{dt} &= p1 + p2_{max}\frac{I(t)^n}{K_m^n + I(t)^n} - d_{MDM2}MDM2(t), \\
\frac{dI(t)}{dt} &= p53(t) \times activity(t) - k_{delay}I(t).
\end{aligned}$$

$degradation(t)$ measures the rate of degradation per bond made between MDM2 and p53 and depends on the stress signal,

$$degradation(t) = degradation_{basal} - [k_{deg} \times signal(t) - threshold(t)],$$

where k_{deg} signifies the repressive effect that a stress signal has on the ability of MDM2 to degrade p53. $threshold(t)$ represents the damping of this repression caused by an

[1]There are various other models that include p53 (Fussenegger *et al.*, 2000; Aguda, 1999; Mao *et al.*, 2001, 1998), but as they are either not at the molecular level or only include p53 as a minor component, they will not be examined here.

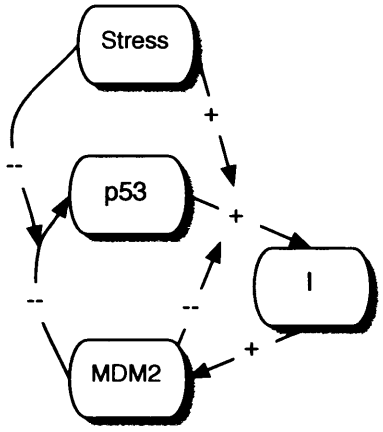


Figure 2.2: A diagram to summarise the model proposed by Bar-Or *et al.*. The stress signal (Stress) disrupts the p53/MDM2 negative feedback loop. The hypothetical intermediary substance (I) delays the effect of p53 on MDM2. This is based upon a diagram in Bar-Or *et al.* (2000).

assumed delay between the delivery of the signal and its effect on MDM2,

$$\frac{d(threshold(t))}{dt} = -k_{damp} \times threshold(t) \times signal(t = 0),$$

The damage signal is assumed to exponentially decay as the DNA is repaired. $activity(t)$ is a quantification of the inhibiting effect that MDM2 has on p53 transcription through MDM2 binding to p53's transcriptional activation region.

The behaviour of the model was observed over a range of parameter estimates. The key finding was that under certain conditions p53 and MDM2 undergo damped oscillations. These oscillations depend on various factors, most significantly the time delay ($k_{delay} > 0$) and the signal. Oscillations are only observed when there is a large initial damage signal. Bar-Or *et al.* went on to find that these findings are in agreement with practical observations. Oscillations were observed, even though they were not measured for more than one cycle and they only occurred at high levels of damage.

It is commendable that the model agrees with practical results but there are weaknesses. Introducing this hypothetical intermediary I seems a sure way to force oscillations into the model but its inclusion is not really justified. Even though the authors claim the model is simple because it has only a few components, it is made complex by the relationships between the components, in all there are 13 parameters that have to be estimated for only three components. This seems to be an unnecessarily large amount. It would be more productive to simplify the relationships but have more components in the p53 network. It seems likely that with an increased number of components, oscillations would still be observed but without the intermediary component.

2.4.2 p53 delay equations

Monk (2003a) explains the usefulness of transcriptional time delays in gene regulatory networks that produce oscillations. The p53/MDM2 system is used as one example. As mentioned above the Bar-Or *et al.* model has a fundamental failing in that it needs a hypothetical intermediary component to produce the seen oscillations. Monk overcomes this by using transcriptional delays. The p53-MDM2 feedback loop is described as,

$$\begin{aligned}
\frac{d[Pp]}{dt} &= \alpha_p - \left(\mu_{p1} + \mu_{p2}\frac{[Mp]^2}{M_0^2 + [Mp]^2}\right)[Pp], \\
\frac{d[Mn]}{dt} &= \alpha_{m0} + \alpha_{m1}\frac{([Pp](y-\tau))^n}{P_0^n + ([Pp](t-\tau))^n} - \mu_{m1}[Mm], \\
\frac{d[Mp]}{dt} &= \alpha_{m2}[Mm] - \mu_{m2}[Mp],
\end{aligned}$$

where $[Pp]$, $[Mm]$ and $[Mp]$ are the concentrations of p53 protein, MDM2 mRNA and MDM2 protein respectively. α_* are production rates, μ_* are degradation rates and τ is the transcriptional time delay. Monk (2003b) finds that when appropriate parameters are chosen "a typical simulation is in good agreement with experimental data". Transcriptional delays are one possible mechanism to create oscillations in the p53 network. It was not clear from the model how cell stress would affect the system, this is important as experiments suggest that at equilibrium oscillations do not occur and only at certain amounts of damage are oscillations seen.

2.4.3 p53 and MDM2 appear to pulse in individual cells

In an interesting study Lahav *et al.* (2004) examined the p53 DNA damage network at the single cell level. Most biological measurements are made on a large population of cells, which gives the average dynamics. This is beneficial as it is relatively easy to make the measurements and has the effect of reducing the biological error in the data. The results are only reliable if the average is a representative dynamic of all the cells. When the dynamics of the cells are radically different information about the dynamics is lost. One approach to dealing with this is to develop techniques to extract single cell data from population data based on models, this is a challenging problem that has not been addressed. The alternative is to make single cell measurements.

Lahav *et al.* developed stable cell lines (MCF7) that expressed fluorescent tagged p53 and MDM2. A population of these cells was exposed to 5Gy γ radiation and then the levels of p53-F and MDM2-F were measured in individual cells using time-lapse fluorescence microscopy. It was found that there were pulses of both p53 and MDM2 with the cell having a range of different numbers of pulses. The duration of the pulses was approximately constant and the MDM2 pulses were out of sync with the p53 pulses by about 100 mins. The mean peak and duration of the pulses remained constant when

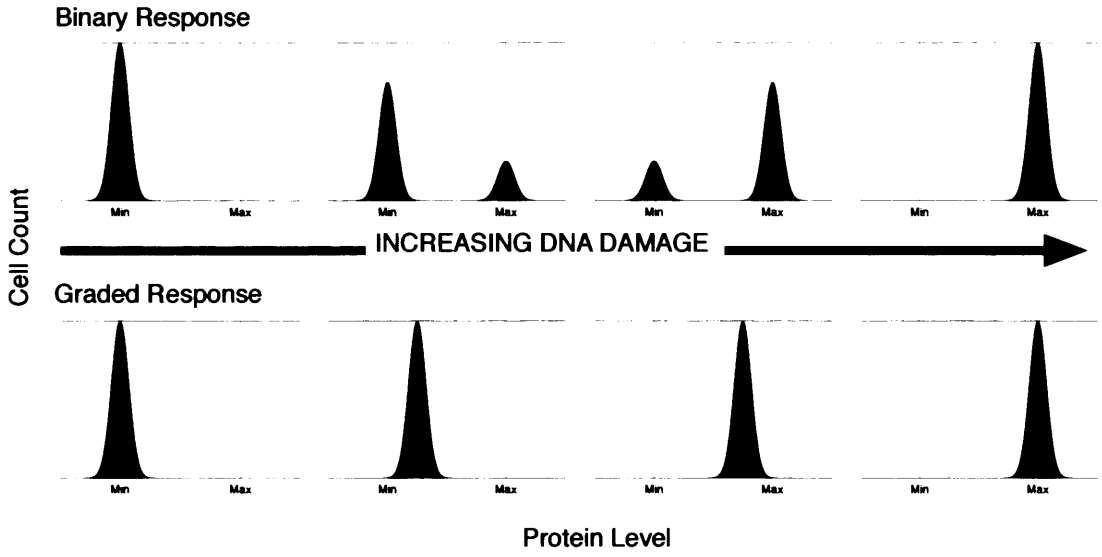


Figure 2.3: Schematic representation of the population distribution of p53 levels for two possible models of response: a binary response and a graded response. Diagram adapted from Jõers *et al.* (2004).

the amount of damage was varied, but the average number of pulses increased. This suggests that sustained oscillations occur between p53 and MDM2 after damage until the cell is repaired or dies.

This is an interesting set of results but more data is required before the possibility that these pulses are damped oscillations can be excluded. An alternative explanation for the pulses could be that the fluorescence is only bright enough to be detected above the background noise when the p53 is tightly packed together i.e. when in the tetramer form in the nucleus. If this was the case then it is indicative that active p53 or the location of p53 changes in pulses.

Another approach to analysing data at the single cell level is to use flow cytometry. Using this approach Jõers *et al.* (2004) examined how the distribution of p53 levels (in NIH 3T3 & MCF7 cells) varied with increasing DNA damage. It was found that levels of p53 for each individual cell increased in a graded way as stress strength increased i.e. as stress is increased the whole distribution of p53 gradually shifts to higher levels (Figure 2.3). This seems to contradict the results of Lahav *et al.*. The results of Jõers *et al.* (2004) are based on a single time point, but if the binary pulse response suggested by Lahav *et al.* holds then the level of p53 should still show a binary response. Interestingly, Jõers *et al.* observe that downstream targets of p53 can show either a binary or a graded response. There are currently no other experiments that confirm these results and until there are, the single cell level results should be treated with some caution.

2.4.4 Models recently published

The possibility of pulse-like dynamics described by Lahav *et al.* (2004) have recently motivated two models of the p53 network that manage to reproduce these dynamics (Ciliberto *et al.*, 2005; Ma *et al.*, 2005). Unfortunately these papers were published as the work on this thesis was near completion and are only described here for completeness.

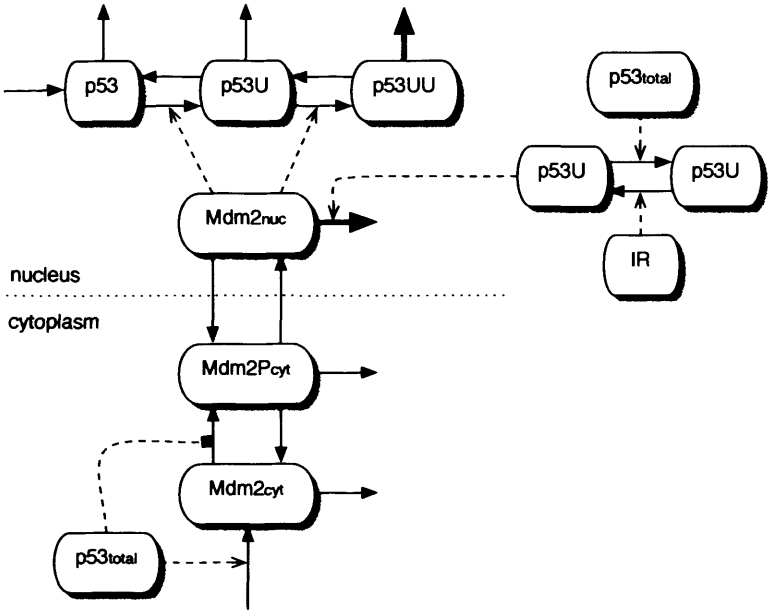


Figure 2.4: A schematic of the p53 model by Ciliberto *et al.*. p53U and p53UU are the two ubiquitinated forms of p53, $Mdm2_{nuc}$ is MDM2 in the nucleus and $Mdm2P_{cyt}$ is MDM2 phosphorylated by AKT that is in the nucleus. Diagram adapted from Ciliberto *et al.* (2005).

Ciliberto *et al.* (2005) have produced an ODE p53 model based at the protein level (Figure 2.4). This model is fairly simple as it only includes MDM2 and p53 as components but there are a number of different forms of each component. p53 is assumed to have a two step ubiquitin process controlled by MDM2 where the rate of degradation increases with each ubiquitin tag added. Of particular importance is the assumption that MDM2 is controlled by its location; it is only allowed to enter the nucleus and hence interact with p53 once it has been phosphorylated by AKT. The rate of this process is negatively regulated by p53 through a rather convoluted loop (see section 2.3) which provides the important positive feedback. DNA damage interacts with the p53/MDM2 system by increasing the degradation rate of MDM2.

Once appropriate parameter values have been chosen the model can replicate the results of Lahav *et al.* (2004). When there is no DNA damage the system remains in a stable steady state, but when gamma radiation is applied, the system shifts to a stable limit cycle. When the DNA damage has been reduced past some critical value, the system

returns to having a steady state equilibrium and eventually returns to its original state. The only area where the model did not agree with the experimental results was that when p53 peaks in the model, MDM2 is very low, whereas in the experimental results MDM2 is already at a high value.

Ma *et al.* (2005) have taken a different approach focusing on DNA damage detection and ATM. They have constructed a mixed framework model with three modules. In the first module, a random number of double-stranded breaks is formed based on a Poisson distribution with an average proportional to the dose. The double-stranded breaks are repaired by a limited number of repair complexes according to a stochastic model. In the second module, the repair complexes cause the activation of ATM, with the rate of activation being a function of the amount of repair complex and active ATM. Active ATM shows switch-like behaviour. The third module is the p53-MDM2 oscillator which is an ODE model whose strong oscillations are driven by both transcriptional and translational delays. If the correct parameters are chosen the solutions replicate accurately the results of Lahav *et al.* (2004). The only exception was that at a high dose (20Gy) the model predicted 100% of cells would have two or more pulses but the experimental data showed only 40% (Lahav *et al.*, 2004). Additionally to replicate the average dynamics of damped oscillations stochasticity has to be introduced in some parameters.

Both of these models are excellent studies that shows that the p53 network can produce pulses when DNA damage occurs. It is interesting that two completely different approaches can produce the same dynamics. The models do have limitations though, as both sets of authors admit themselves, their models are very selective about the mechanisms and components that they include. A limitation of the model of Ma *et al.* (2005) is that it does not include localisation effects. It is also unclear whether all the assumptions hold up to scrutiny and further experimental evidence would be needed to clarify certain results. Overall though, these models lay the basis for future models.

Chapter 3

The experiment and data analysis

3.1 Introduction

The overall goal of this work is to produce a better description of the DNA damage response through building a dynamic mathematical model. This requires measurements of the time course of the DNA damage response at both the protein and mRNA levels. To optimally achieve this, one has to focus on a specific biological system that is well defined and controllable because there is considerable variability in the response depending on the system. Here the DNA damage response of MOLT 4 cells will be examined as it is a well established human cell line with an intact response which means that any discoveries can be applied to medicinal purposes. MOLT 4 cells are also a good experimental system as the cells are easy to grow and can be successfully transfected. A reliable way to cause DNA damage is to expose to the cells to ionising radiation.

Time course data was gathered for mRNA and various proteins after MOLT 4 cells had been exposed to ionising radiation at different doses giving measurements of the average DNA damage response. The protein data was quantified and analysed. It was found that the electrochemical luminescence method used for protein detection on Western blots was difficult to reliably quantify, leading to significant measurement error. An alternative detection method was investigated and found to be an improvement. The amount of double-stranded DNA damage was quantified by counting H2AX foci which indicated sites of DNA repair complex formation in the nucleus.

All experiments were performed by Daniela Tomescu, Mike Hubank or Kai Rothkamm, but the quantification and analysis of the data was performed by the author.

3.2 Materials and methods

3.2.1 Cell line

The cell line used was human MOLT4 cells (T cell acute lymphoblastic leukaemia) obtained from NIBSC, UK (CFARP011). These were shown to have an appropriate and intact DNA damage response (Figure 3.1): the wild type status of p53 was confirmed by sequencing the p53 genotype of both alleles; Western blot experiments demonstrated that p53 accumulation occurred after irradiation and QPCR showed that known p53 targets (p21, GADD45α & MDM2) were activated after ionising radiation. These experiments were performed by Daniela Tomescu.

3.2.2 Experiment

Human MOLT4 cells in log phase (10^6 ml^{-1}) were γ-irradiated with 5 Gy at a dose rate of 2.45 Gy per minute (using a ^{137}Cs γ-irradiator). Cells were harvested at 2 hour intervals up to 20 hours past irradiation. Each sample was then divided into two, one half being used to measure mRNA levels and the other to measure protein levels. This

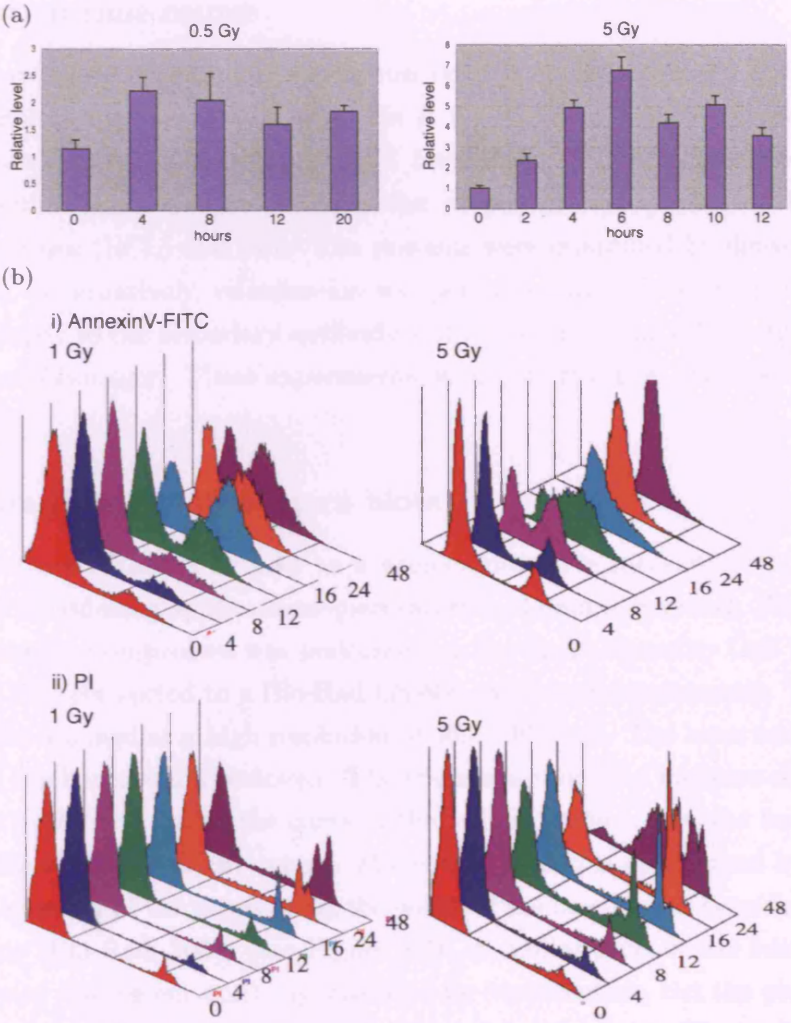


Figure 3.1: MOLT4 DNA damage response validation. (a) QPCR measurements of MDM2 (a known p53 target). (b) Apoptotic marker cell population charts (AnnexinV-FITC is an indicator that the cell is committing apoptosis and PI that the cell is dead).

experiment was run three times on independent cell preparations. The above experiment was repeated once with 0.5Gy of radiation. Cells were harvested at 0, 2, 4, 6, 8, 10, 12, 16 and 20 hours. These experiments were performed by Daniela Tomescu and Mike Hubank.

3.2.3 Microarray time course

For each mRNA sample, RNA and cRNA were prepared, and their quantity and quality determined by Nanodrop spectophotometer and Bioanalyser 2100 (Agilent). Affymetrix microarray measurements were then made on all mRNA samples between 0 and 12 hours post irradiation, using Affymetrix U133A arrays (see appendix A.3). The gene expression levels were then calculated using the Affymetrix MAS5.0 algorithm (Affymetrix, 2002a,b). This was performed by Daniela Tomescu and Mike Hubank.

3.2.4 Protein time course

For each sample, protein was extracted, run on a polyacrylamide gel and transferred to a nitrocellulose membrane (see appendix A.1). Western blots were probed against total p53 (Santa Cruz), phospho-p53 (Cell Signalling), MDM2 (Oncogene and Santa Cruz) and actin (Santa Cruz). Visualisation of protein was performed by enhanced chemiluminescence (ECL) and film. The proteins were quantified by densitometry (see section 3.2.5). Alternatively, visualisation was performed by directly coupling a Cy3 fluorescent antibody to the secondary antibody and measuring using a Molecular Dynamics Typhoon phosphoimager. These experiments were performed by Daniela Tomescu and Mike Hubank.

3.2.5 Quantification of Western blots

Western blots are regarded at best as a semi-quantitative measure of the amount of protein but for modelling applications quantification of data is essential. When ECL was used the quantification process was performed on Bio-Rad's Quantity One 1-D Analysis Software (v4.2.1) connected to a Bio-Rad GS-800 calibrated densitometer. The Western blot films were scanned at a high resolution of 36.3×36.3 μm. The lanes and bands were defined, and the background removed. The trace was used as a measure of the amount of protein; it is the area under the curve of the optical profile where the band is defined and is in units of optical density×mm. The optical profile is determined by calculating the average intensity of the pixels along the width of the lane; this is done for every point along the lane (Bio-Rad, 2005) (see Figure 3.2). A similar process was followed when a directly coupled fluorescent antibody was used for visualisation, but the phosphoimager directly provided an image and this was manipulated using ImageQuant v5.1 software.

The raw protein value has to be adjusted to correct for error and allow the results to be comparable. The amount of protein is defined as x_i where i is the lane number. The first process was to normalise between the lanes; although the lanes are loaded as equally as possible there is still some variation in the total protein per lane. This was done by measuring, concurrently with the actual protein measurement, the amount of a control protein that is assumed to remain constant whatever happens to the cell; commonly actin is used. Values for the amount of actin (a_i) were obtained using the same process as described above. The protein values were then adjusted to the values they would have if the loading was balanced i.e. $x_{actin,i} = \frac{x_i}{a_i/a_0}$ (a_0 is the first lane's actin value).

There are a large number of factors, such as the film exposure time and amount of total protein, that can vary from Western blot to blot. This variability was taken account of by placing a standard on every blot. The standard was taken from a large sample of cells exposed to 5Gy of radiation and frozen down in aliquots at 6 hours. The measured values of protein are converted to units of standard i.e. $x_{final,i} = \frac{x_{actin,i}}{x^*}$, where $x^* = x_{actin,j}$ and j indicates the lane of the standard. This only allows the comparison of

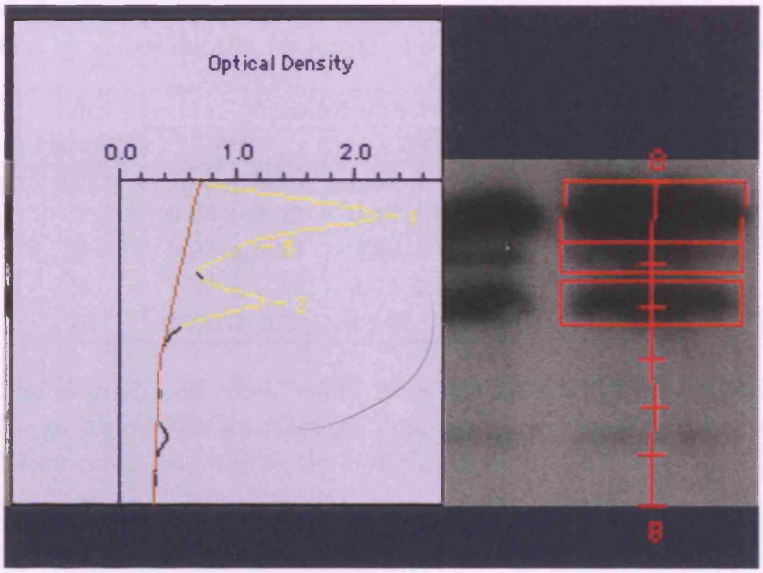


Figure 3.2: An example of the profile generated from one lane using Quantity One 1-D Analysis Software. The yellow parts indicate where bands are defined. The brown line indicates the background that is removed.

the *same* protein in different situations. To compare different proteins numerically one would need to assign absolute values by quantifying the absolute amount of standard for each protein. In summary the data processing is defined as follows:

$$x_{final,i} = \frac{x_i}{a_i/a_0} \Big/ x^*$$

where x^* is the actin adjusted standard.

3.2.6 H2AX

γ-H2AX (a fluorescent antibody specific for the phosphorylated form of H2AX) was used to count the number of double-strand breaks in MOLT4 cells at 0.5, 1, 2, 4 and 17 hours after exposure to 0.05, 0.2 or 0.5 Gy of γ radiation. This was performed by Daniela Tomescu and Kai Rothkamm using the protocol in Kühne *et al.* (2004).

3.3 The time course of double-stranded DNA breaks

The input to the p53 network is the number of DNA breaks. It is possible to get a measure of this using a fluorescent antibody specific for the phosphorylated form of H2AX (γ-H2AX). Early after a double-stranded break is formed H2AX is phosphorylated on serine 139 and moves close to the break (Rogakou *et al.*, 1998). It has been established that the number of foci observed when γ-H2AX is applied is equivalent to the number of double-stranded breaks (Sedelnikova *et al.*, 2002; Rothkamm and Löbrich, 2003). There-

Table 3.1: The average number of breaks per cell after irradiation with γ rays. The count is adjusted by removing the background count (0.380 ± 0.093 breaks).

Time (hours)	Amount of γ radiation (mGy)		
	500	200	50
0.5	18.2 ± 0.56	6.88 ± 0.39	1.84 ± 0.20
1	9.64 ± 0.49	4.42 ± 0.27	1.51 ± 0.18
2	5.22 ± 0.29	2.89 ± 0.24	0.786 ± 0.16
4	2.40 ± 0.35	1.14 ± 0.19	0.321 ± 0.16
17	0.397 ± 0.21	0.142 ± 0.16	-0.0556 ± 0.12

Table 3.2: The degradation rates found from the H2AX data with and without the final point. The p-value is the probability that the linear regression is a correct model assuming Gaussian error (the higher the better, see section 8.2.2).

Amount of γ radiation (mGy)	with 17 hour point		without 17 hour point	
	Degradation rate (hr^{-1})	p-value	Degradation rate (hr^{-1})	p-value
500	0.181 ± 0.018	0.126	0.510 ± 0.15	0.683
200	0.200 ± 0.014	0.0131	0.478 ± 0.087	0.903
50	0.752 ± 0.0097	0.00131	0.505 ± 0.064	0.794

fore γ-H2AX is a useful marker of a single break. γ-H2AX was used to count the number of double-strand breaks in MOLT4 cells at 0.5, 1, 2, 4 and 17 hours after exposure to 0.05, 0.2 or 0.5 Gy of γ radiation. At each time point the breaks in 50 to 100 cells were counted (Table 3.1).

DNA damage repair is modelled assuming repair occurs at a constant rate. This means that the number of breaks would be expected to decline exponentially. To find the rate of repair, the data is converted to the log domain and a linear regression is performed using the weighted least squares error function (see section 8.2.2). The error is transformed into the log-domain using the assumption that the data is distributed as a log normal and,

$$s^2 = \ln\left(\frac{\sigma^2}{\mu^2} + 1\right),$$

where s is the error in the log domain, σ is the error and μ is the average in the linear domain (this is only an estimate as the assumption is incorrect) (Rice, 1995). The linear fit was performed with and without the last point.

When the 17 hour time point is included in the plot the p-value is considerably worse than without it (Table 3.2). Also, at latter points the rate of repair slows down more than one would expect if the repair was exponential (Figure 3.3). This could simply be because of measurement error and the number of breaks is at the background level; the average number of breaks at 17 hours are generally in agreement with zero breaks. Further work in this group on the distribution of breaks suggests that something more complex is occurring (Barenco, personal communication) and that exponential repair may not be

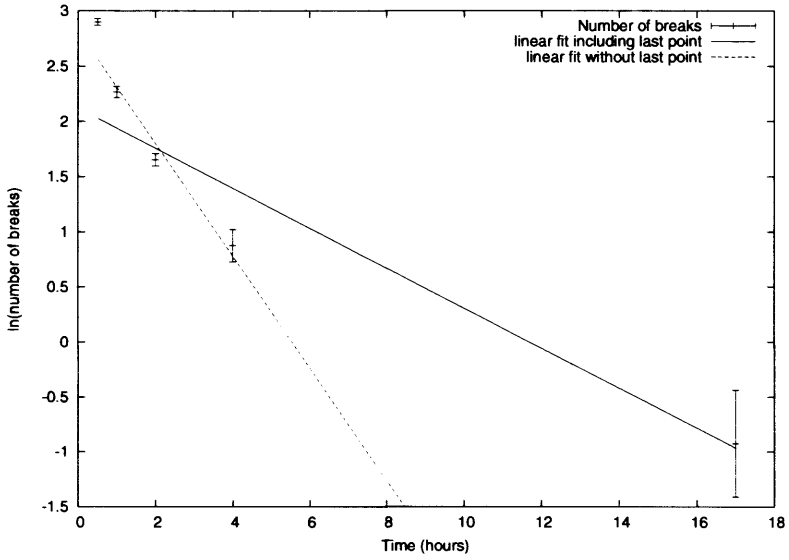


Figure 3.3: A plot showing the H2AX data in the log domain when the cells are exposed to 0.5Gy of ionising radiation. The linear fits both with and without the last data point are shown.

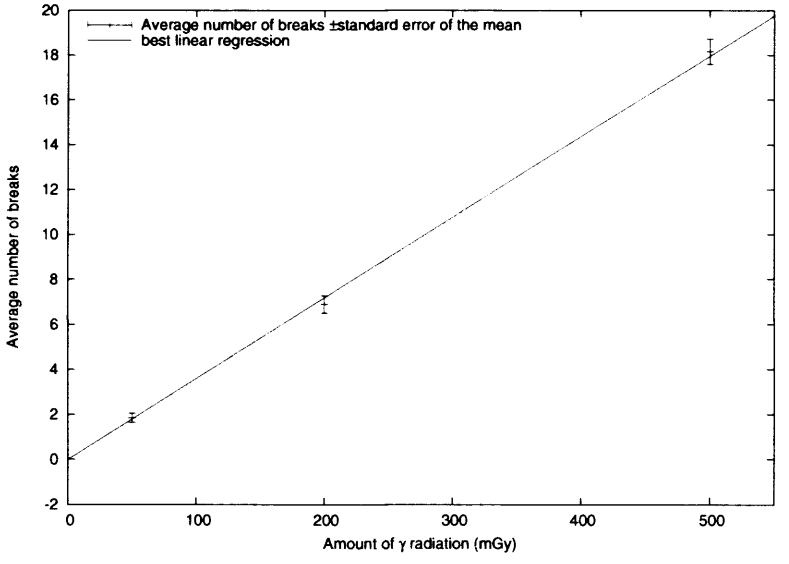


Figure 3.4: A plot showing that the initial damage caused by ionising radiation is proportional to the amount of radiation. The gradient of the linear regression is 0.036 and the y-intercept it -0.0053.

the best description of DNA repair. There could be two biological explanations for this (Barenco, personal communication),

1. The DNA damage is not homogeneous with some damage harder to repair than other damage. The harder to repair damage would have a lower rate of repair hence as the easy to repair breaks are mended, the overall repair rate slows. This has also been suggested by Rothkamm and Löbrich (2003).

2. That there is some feedback in the system that introduces new breaks in the DNA.

Due to the uncertainty to the causes of this apparent slow down in repair and to keep simplicity in the models, it was decided to omit the 17 hour time point.

The predicted degradation rates when the last point is not included are all in agreement (Table 3.2). This is an interesting result and suggests that the rate of repair is constant whatever the damage. In this thesis the repair rate will be considered to be 0.5 hr^{-1}, which is equivalent to a half life of 1.4 hours. Another significant finding is that the initial number of breaks (at 0.5 hours) appears to be proportional to the amount of radiation (Figure 3.4). This makes it reasonable to extend the findings made here to larger amounts of damage. The linear regression is a good fit and the y-intercept is close to zero (-0.0053) as would be expected if this was a linear process. Another study that used a much broader range of radiation doses also found that there was a linear relationship between the initial number of breaks and the amount of radiation, with 35 double-stranded breaks per cell per Gy (Rothkamm and Löbrich, 2003). This was using human MRC-5 cells, and is very similar to the 36 double-stranded breaks per cell per Gy observed here.

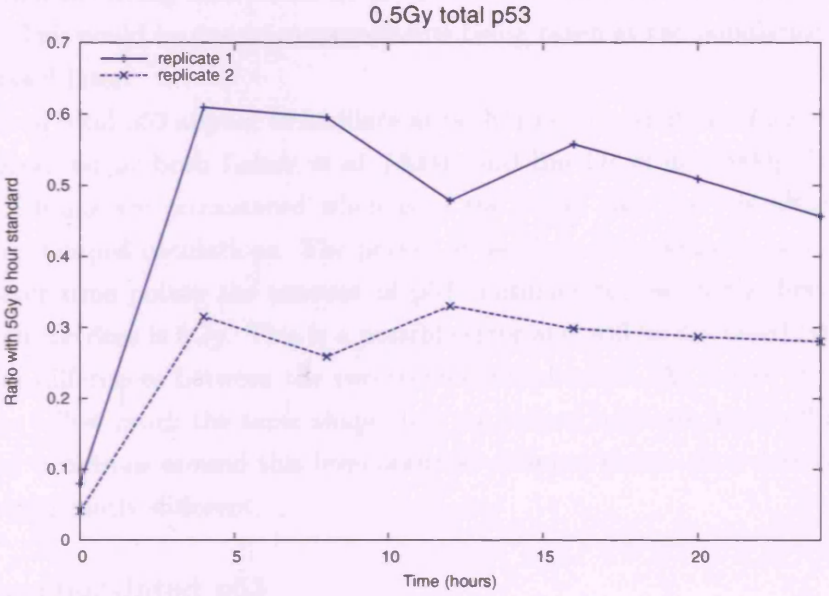
3.4 Protein results

MOLT4 cells containing functional p53 were grown and irradiated with two strengths of radiation (0.5Gy and 5Gy). Protein and RNA were extracted at regular intervals after irradiation. mRNA and protein time courses of the response to DNA damage were produced by microarray experiment and Western blot. Three biological replicates were performed at 5Gy and one at 0.5Gy.

3.4.1 p53

After irradiation the protein levels of *total* p53 (active and inactive) initially rise (Figure 3.5). This is expected as when a cell feels stress the p53/MDM2 negative feedback loop that normally suppresses the amount of p53 is disrupted, allowing the levels of p53 to rise. The time course is different depending on the level of radiation; at 0.5Gy the amount of total p53 levels off after 4 hours whereas at 5Gy the amount rises at a quicker rate and for a longer duration. This is because the damage signal is stronger at higher

(a)



(b)

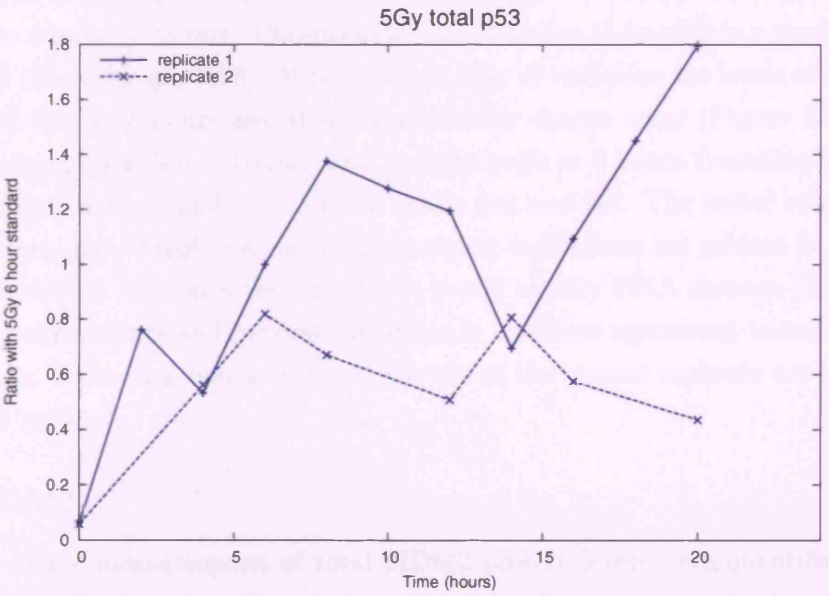


Figure 3.5: A comparison of the total p53 level for two *technical* replicates (two measurements on the same experimental sample) when the level of radiation was (a) 0.5Gy and (b) 5Gy. In replicate 1 the standard was not available so the 5Gy 6 hour time point was used as the standard.

doses. The initial levels are similar at both doses suggesting the protein quantification is working correctly. At the later time points the levels of protein are distant from the pre-damage level. One would expect that after 20 hours the DNA damage signal would have ceased and the living cells would be in their unstressed state but the data does not suggest this. This could be due to measurements being taken at the population level and will be discussed later.

The levels of total p53 appear to oscillate at both doses of radiation. This agrees with the results observed by both Lahav *et al.* (2004) and Bar-Or *et al.* (2000). In the 5Gy case the oscillations are pronounced whereas in the 0.5Gy case there is some evidence that there are damped oscillations. The period of oscillations appears to be variable.

At the later time points the amount of p53 continues to rise in the first technical replicate when the dose is 5Gy. This is a possible error and will be discussed later. There are significant differences between the two technical replicates[1]. At a dose of 0.5Gy, the two replicates follow much the same shape, but the second replicate levels off at a lower value and the variations around this level occur at different times. At a dose of 5Gy the shapes are significantly different.

3.4.2 Phosphorylated p53

The functionally active form of p53 is expected to have different dynamics to the total amount of p53 because DNA damage causes active p53 to become the dominant form when prior to damage it is rare. Phosphorylated p53 at Ser 15 (p-p53) is a good indicator of active p53 (Banin *et al.*, 1998). When there is 5Gy of radiation the levels of active p53 rise about 12 fold in 4 hours and then exponentially decays away (Figure 3.6). In the 0.5Gy case the pulse is less extreme, with a sharp peak at 4 hours (reaching about four times the initial amount) and then a more gentle rise and fall. The initial values for the two doses are similar. There is some indication that oscillations are present in the 0.5Gy case. Two technical replicates were made for p-p53 at 5Gy DNA damage (Figure 3.7). The first few time points and general shape are in excellent agreement between the two measurements. From four hours on the the levels of the second replicate are lower than in the initial replicate.

3.4.3 MDM2

Three time course measurements of total MDM2 protein levels were quantified at both doses of radiation. When the cells are exposed to 5Gy, the general trend is for the amount of MDM2 to gradually decrease over 20 hours (see Figure 3.8(a)). At 8 hours there appears to be a brief increase (especially for replicate 3) before continuing downward. The amount of MDM2 is reduced by about half in 20 hours. When cells are irradiated by

[1]Two measurements on the same experimental sample.

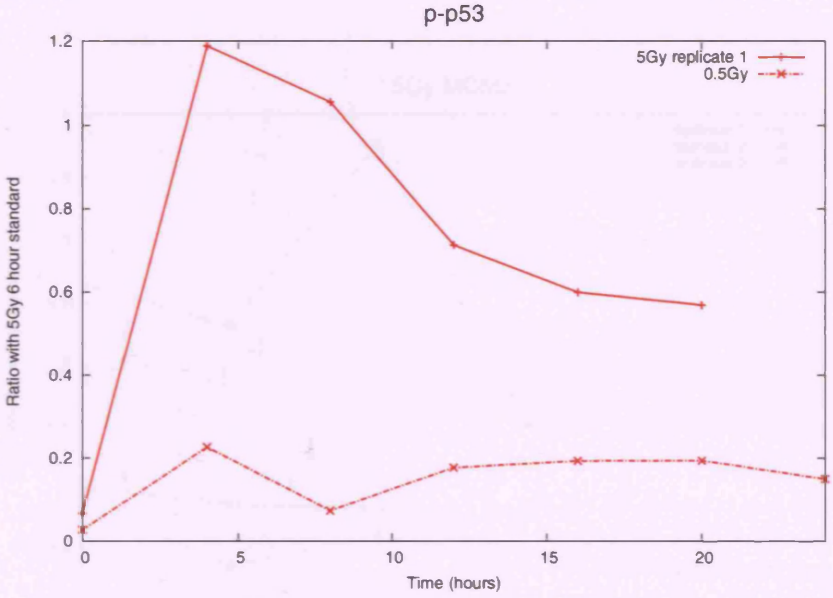


Figure 3.6: A plot showing how the levels of phosphorylated p53 (p-p53) vary after the cells have been irradiated (at levels of 0.5Gy and 5Gy).

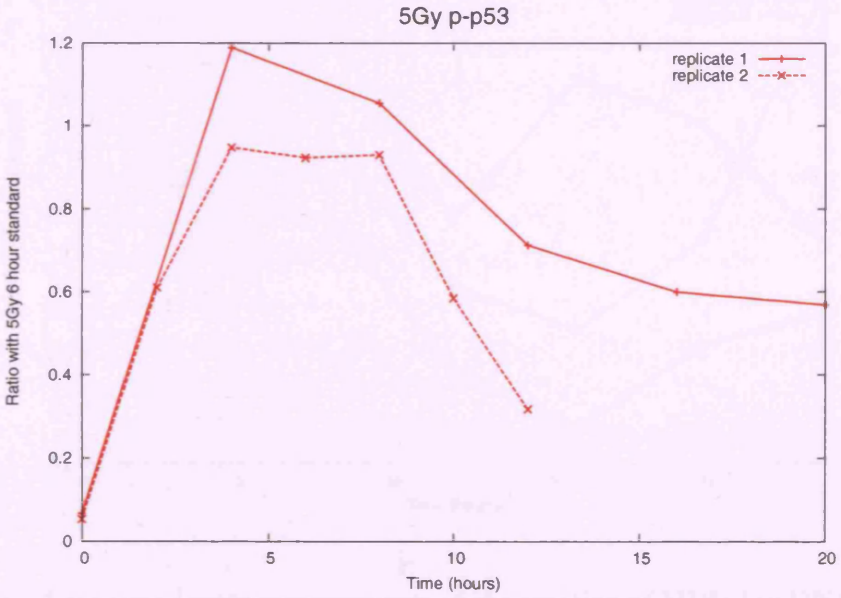
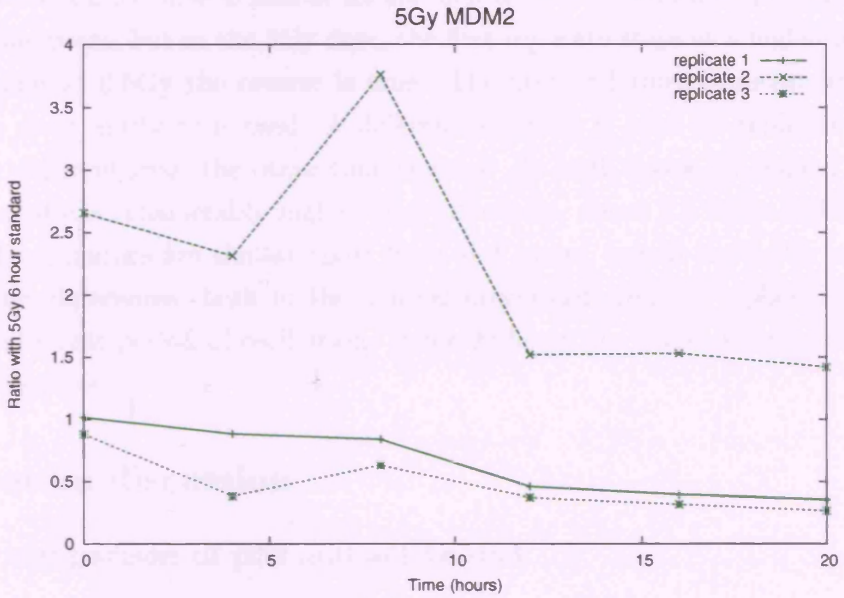


Figure 3.7: A plot showing how the levels of phosphorylated p53 (p-p53) vary after the cells have been exposed to 5Gy of radiation for two different technical repeats of the measurements.

(a)



(b)

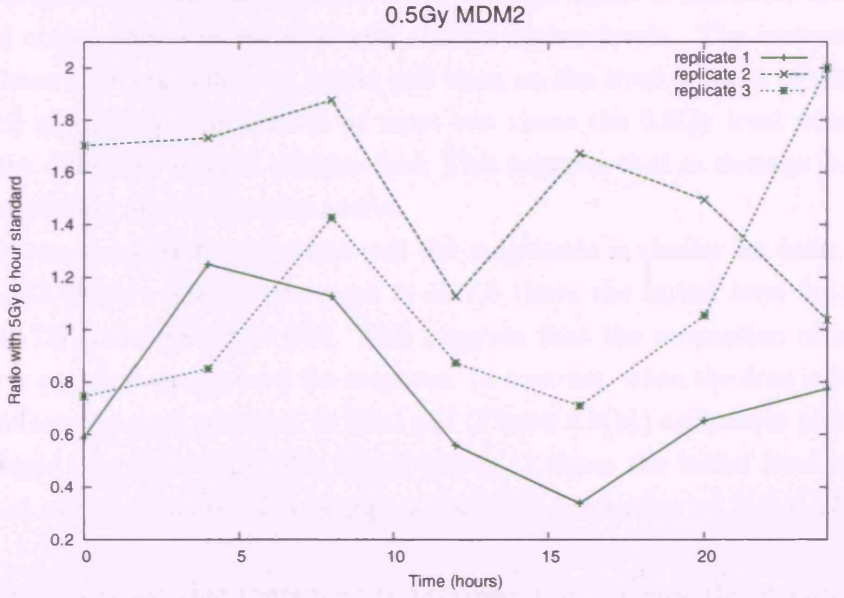


Figure 3.8: Three time course measurements of the reaction of MDM2 to DNA damage caused by (a) 5Gy and (b) 0.5Gy of radiation. Replicate 1 and 3 use the Oncogene antibody whilst replicate 2 uses the Santa Cruz antibody. All protein used was taken from the same biological sample.

0.5Gy the response is different from the 5Gy case (Figure 3.8(b)); there are oscillations that occur around the initial protein level.

There are considerable differences between the three time courses. At both doses, the shape of the time course is similar for the first and third replicate apart from at the four hour time point, but at the 5Gy dose, the first replicate stays at a higher level than replicate 3 and at 0.5Gy the reverse is true. The first and third replicate are similar because the same antibody is used. A different antibody is used for replicate 2 and it is markedly different from the other time courses. At both doses the second replicate measurement starts considerably higher than the others, about 2.5 times higher. In the 5Gy case the dynamics are similar apart from at 8 hours, but in the 0.5Gy case there are significant differences, both in the general downward trend of replicate 2 and an apparently different period of oscillation. After 20 hours the levels seem to diverge for all the time courses.

3.5 Protein discussion

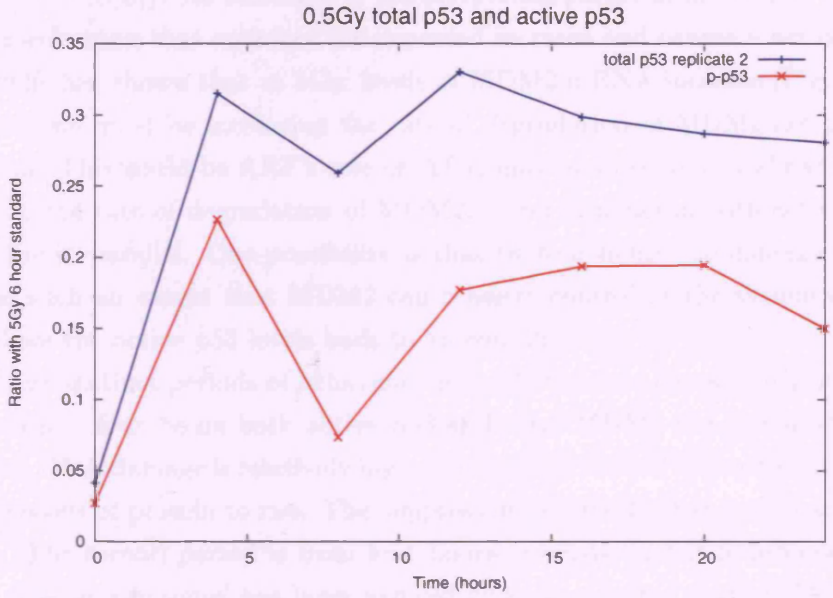
3.5.1 A comparison of p53 and active p53

The general response of a cell to DNA damage is to allow the levels of p53 to rapidly rise. This is true for the MOLT 4 cells with both active and total p53 rapidly rising at both doses (Figure 3.9). As the strength of the damage signal is increased the response of total and active p53 is more rapid and reaches higher levels. The increase in dose appears to have a greater effect on active p53 than on the total amount of p53; for the total amount of p53 the 5Gy level is at most two times the 0.5Gy level whereas with active p53 the difference reaches thirteen fold. This suggests that as damage is increased a greater proportion of p53 becomes active.

At 0.5Gy the shape of the response and the magnitude is similar for both total p53 and active p53 (Figure 3.9(a)); the peak is at 7.0 times the initial level for total p53 whereas it is 7.3 times for active p53. This suggests that the proportion of active p53 remains fairly constant throughout the response. In contrast, when the dose is 5Gy active p53 has a different shaped response to total p53 (Figure 3.9(b)) and active p53 increases at a more rapid rate than total p53; active p53 is 17 times the initial level at 4 hours whereas total p53 is 9 times. This suggests that the proportion of active p53 rapidly increases.

These results suggest that there are two processes that increase the amount of active p53 in the initial DNA damage response: the total amount of p53 is increased, and the proportion of p53 that is active is increased. At medium levels of damage the rise in the total amount of p53 is the dominant process whereas at high levels the proportion of active p53 plays a role. There may be independent mechanisms regulating these two processes.

(a)



(b)

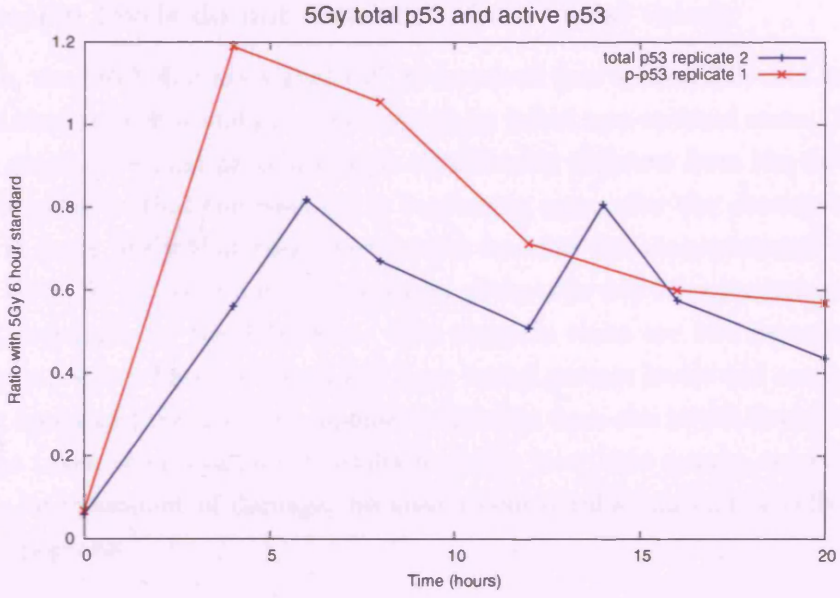


Figure 3.9: A plot comparing the response to radiation of total p53 (p53) and phosphorylated p53 (p-p53) levels after (a) 0.5Gy and (b) 5Gy of DNA damage. The values are not directly comparable. The first replicate is displayed.

3.5.2 A comparison of MDM2 and active p53

The key interaction in the p53 DNA damage network is between MDM2 and p53. At a dose of 5Gy there is a large increase in active p53 but no corresponding increase in MDM2 (Figure 3.10(b)). As MDM2 is a transcriptional target of active p53 there must be another mechanism that counters the expected increase and causes a net decrease in MDM2. QPCR has shown that at 5Gy, levels of MDM2 mRNA increase (Figure 3.1(a)) so the mechanism must be increasing the rate of degradation of MDM2 not preventing its production. This could be ARF's role or ATM may be directly or indirectly causing an increase in the rate of degradation of MDM2. After four hours both active p53 and MDM2 decline in parallel. One possibility is that by four hours the damage signal has decreased to such an extent that MDM2 can reassert control of the system and hence begin to reduce the active p53 levels back to an equilibrium.

There is two distinct periods of behaviour in the 0.5Gy time courses (Figure 3.10(a)). Between zero and four hours both active p53 and total MDM2 are increased and it is proposed that DNA damage is relatively high and the p53/MDM2 interaction is disrupted allowing the levels of protein to rise. The suppression of MDM2 that occurs at 5Gy does not appear. The second period is from four hours onwards, and it is proposed that in this period the damage signal has been reduced by such a degree that MDM2 begins to act as a regulator of p53 again, producing oscillations.

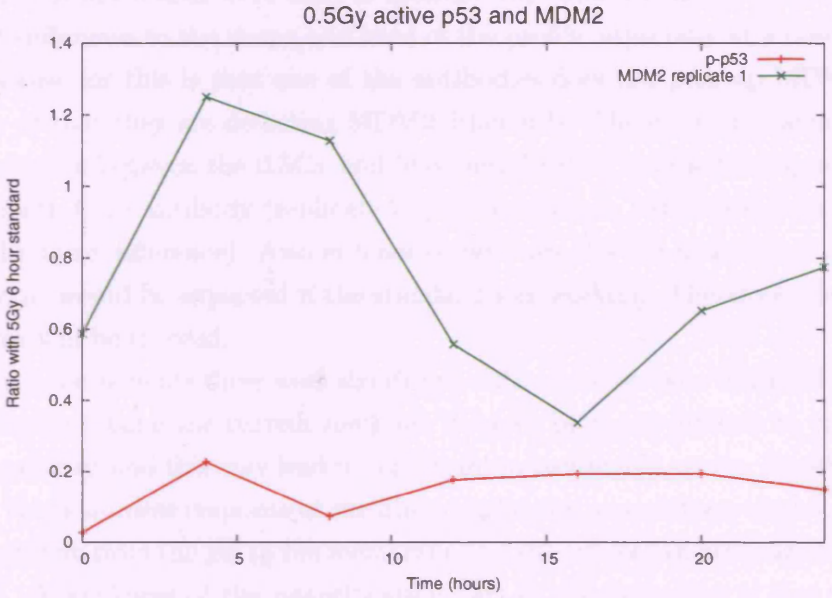
3.5.3 Protein levels do not return to their initial values

By 20 hours, the DNA damage signal will have ceased (see section 3.3) and so it would be expected that the cell would have returned to its initial non-stressed state. This is not seen; in all results the final protein level is significantly different from the initial value. One possible cause is that the response is continuing even after the damage signal has ceased. It is more likely that this effect occurs because the measurements are at the population level. At 20 hours a high proportion of the cells will be committing apoptosis or be dead, especially at the 5Gy dose. This suggests there are two separate populations, one group which have recovered and have initial protein levels and another that is committing apoptosis; the average response is different from the initial levels. Generally, care must be taken when examining results from the later time points, especially where there was a large amount of damage, because a considerable amount of cells will have committed apoptosis.

3.5.4 Quality of data collection

The first replicate time course for total p53 at a dose of 5Gy continues to rise from 14 hours to 20 hours (Figure 3.5(b)). This does not seem correct as one would expect that by 20 hours most cells would be dead or dying, which implies that the levels should not be rising. On the Western blot x-ray most of the bands were very blurry and over-

(a)



(b)

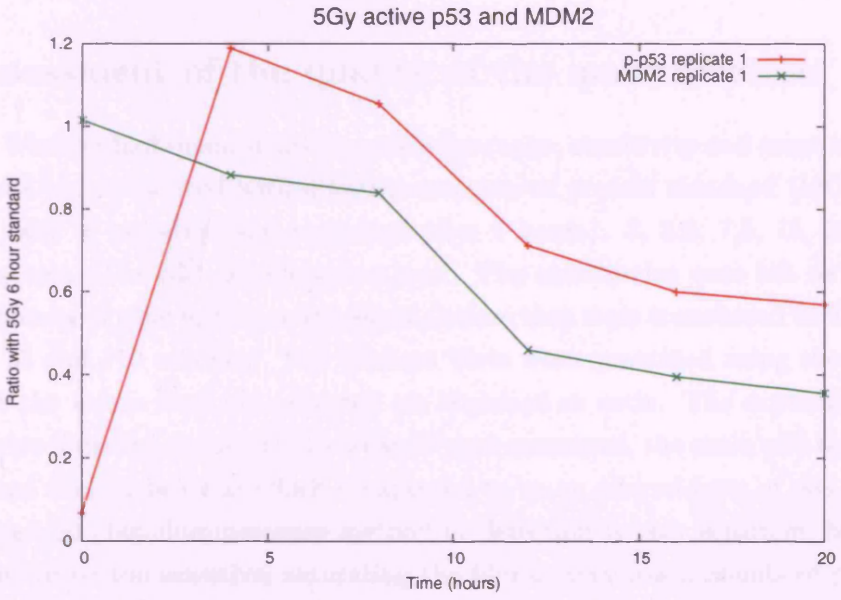


Figure 3.10: A time course plot comparing the response of phosphorylated p53 and MDM2 (first replicate) at (a) 0.5Gy and (b) 5Gy of radiation. The values of the different proteins are not directly comparable. MDM2 is the total amount of MDM2 both active and inactive (in regard to the interaction with p53).

exposed and the corresponding actin blot had a high background and lots of bubbles. Even without the actin adjustment the level of p53 still rises. This data should be treated with caution.

Two different antibodies were used to measure the total MDM2 level and this made a significant difference to the shape and level of the profile, especially at a dose of 0.5Gy. A possible cause for this is that one of the antibodies does not pick up MDM2 as well as the other or that they are detecting MDM2 differently. There are only approximately 0.5 standard units between the 0.5Gy and 5Gy initial values for the Oncogene antibody whilst the Santa Cruz antibody (replicate 2) produces about 1 standard units difference (there should be no difference). Also at 6 hours, replicate 2 is not near a level of 1 after a dose of 5Gy, as would be expected if the standard was working. Therefore, the replicate 2 time course will be ignored.

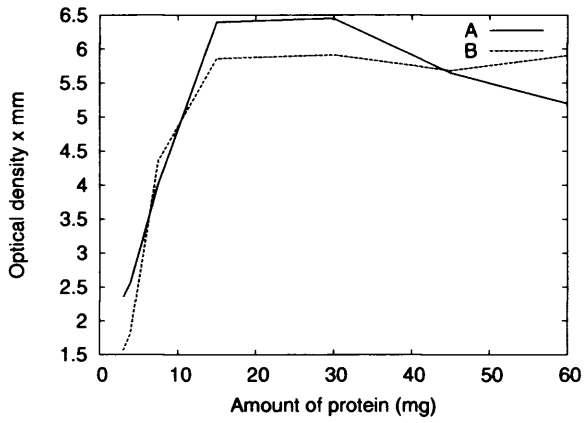
In all three components there were significant differences between technical replicates. This suggests that using the current methods Western blots are difficult to reliably and accurately quantify, and this may lead to significant measurement errors. Possible sources of error are the non-linear response of the film to light, the non-uniform electric field used to transfer protein from the gel to the membrane and the subjective defining of bands and background. A weakness of the quantification part of the procedure is that the values are very sensitive to the standard, the errors in quantifying the standard are propagated to the rest of the measurements.

3.6 Assessment of the quality of the quantification

To test the Western blot quantification process for range, sensitivity and amount of error, Western blots were produced with different amounts of protein standard (MOLT4 cells exposed to 5Gy of radiation and harvested after 6 hours): 3, 3.9, 7.5, 15, 30, 45 and 60 mg of protein. The p53 antibody was used. The membranes were left to decay for different amounts of time to reduce the signal, before they were transferred to film: 5, 10, 15, 30, 50, 75 and 110 minutes. The Western blots were quantified using the standard method but the values were not adjusted *via* standard or actin. The experiments were repeated twice (labelled A and B). Two bands were measured, the main p53 band (band 1) and a band slightly below it which is expected to be an altered form of p53 (band 2).

The enhanced chemiluminescence method of detection is very sensitive, but in this case it seems to be too sensitive, saturating the film at very low amounts of protein no matter how short a period the film is exposed (Figure 3.11). For band 1 the saturation occurs at around 15 mg of protein whereas for band 2, the linear region is larger (ending between 15 and 30 mg) because there is less protein. When the gel is only left to decay for a short time the optical density remains high after saturation (Figure 3.11(a)) but when there is a longer gel decay time the optical density drops after saturating (Figure 3.11(b)). This may be because at a higher concentration of protein the signal

(a)



(b)

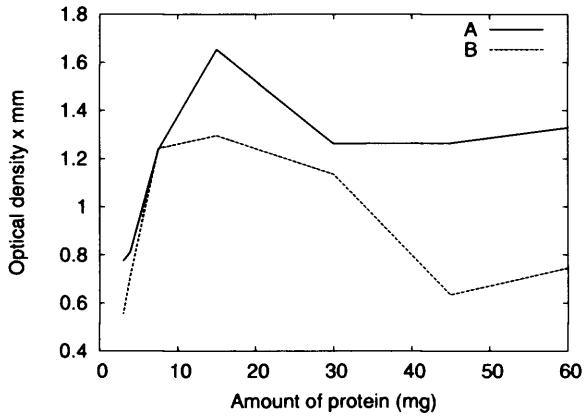


Figure 3.11: Example plots of how the optical density of the main p53 band varies with concentration for the two replicates A and B. This is when the decay time is (a) 5 and (b) 110 mins.

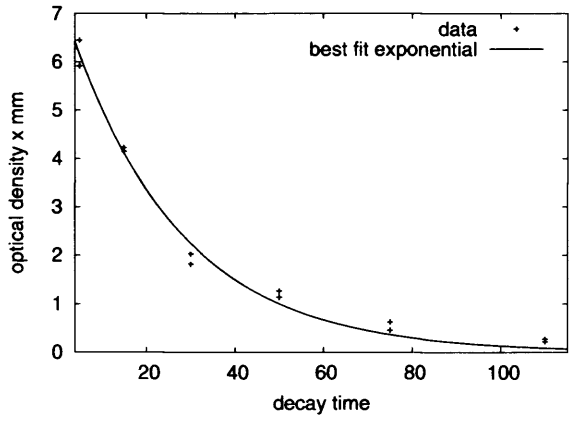


Figure 3.12: An example of how the signal decays as the time before fixing increases. This is from data with 30 mg of protein (Band 1).

degrades at an increased rate. Some bands on the film that have been left to decay for a long time have a white central region which supports this hypothesis.

As the blots are left to decay the relative amount of protein between bands does not appear to be preserved. Even though the signal degrades with decay time in an exponential manner (Figure 3.12), the degradation rate increases with increasing amount of protein (Figure 3.13). This suggests that leaving the Western blot to decay seriously affects the quantification.

As the duration of decay increases the relative error between the two technical replicates averaged over the amount of protein also increases (Figure 3.14(b)). The errors are large ranging from 15 to 40%. Also, at high levels of protein the average relative error is generally larger (Figure 3.14(a)). This is because at high levels the film is more likely to saturate which is likely to produce greater errors in defining the bands. The error for Band 2 is generally lower than Band 1, because the signal is lower and so is less prone to the non-linear effects of saturation.

These results suggest that to get a good grasp of what is occurring at the protein level after DNA damage many replicates are required at a large number of time points. Most of the time this will not be practical. An alternative is to adapt the Western blot quantifying procedure to remove steps in the process and hence reduce the error. One such approach is considered below.

3.7 An alternative way of visualising Western blots

In an attempt to improve the quality of the Western blot data it was decided to use a different detection method and hence a different secondary antibody. In this method the secondary antibody is directly coupled to a fluorescent antibody, Cy3. Light is emitted when light at a certain frequency is shone on the membrane and the resulting fluorescence can then be measured directly by a phosphoimager. This improves on the previous procedure by removing the need to have film, the phosphoimager directly scans the membrane. This will remove the source of error that comes from using film, so this procedure should produce more reliable results. This technique is less sensitive than the previous procedure, but as there is plenty of material available this is not a concern. The resulting images are quantified in a very similar way to the standard Western blot method (see section 3.2.5). Unfortunately, due to a lack of resources and other problems the actin gels could not be used, so it had to be assumed that an equal amount of total protein was applied to each lane.

The images produced by this method are a lot cleaner, with less background signal and sharp rectangular bands (Figure 3.15). For total p53, the time course for both doses start at the same initial value and rise rapidly (Figure 3.16). Total p53 rises faster when there is 5Gy of damage and peaks twice at 6 and 14 hours, before declining. When there is 0.5Gy of damage the increase is slower and it peaks at a lower value, it peaks twice at 8 and

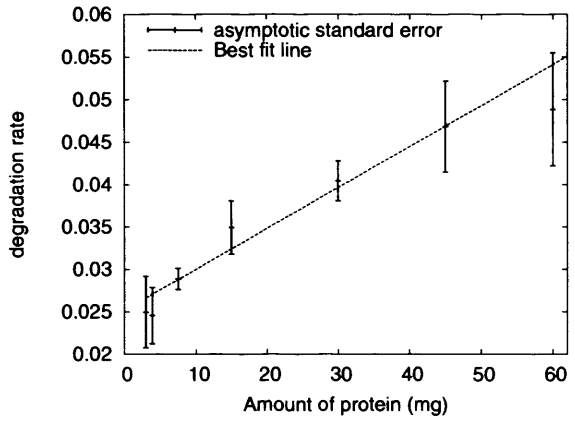
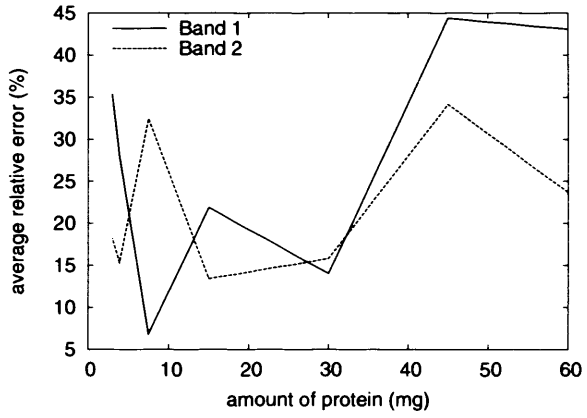


Figure 3.13: A plot showing how the degradation rate of the signal increases as the amount of protein increases. This is for the main p53 band.

(a)



(b)

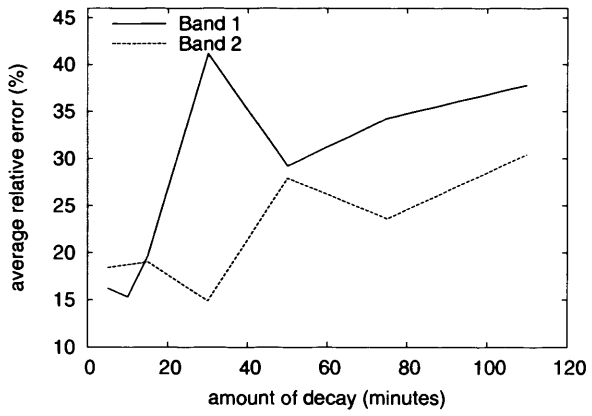


Figure 3.14: A plot of showing how the relative error of both p53 bands varies with (a) the amount of protein and (b) the duration of decay.

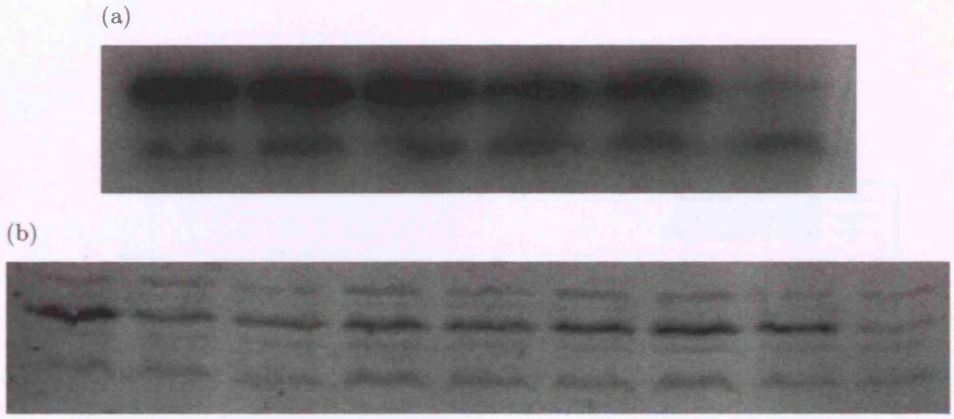


Figure 3.15: The images produced for p53 after 5Gy damage by (a) the standard Western blot procedure and (b) the adapted procedure using the Typhoon phosphoimager.

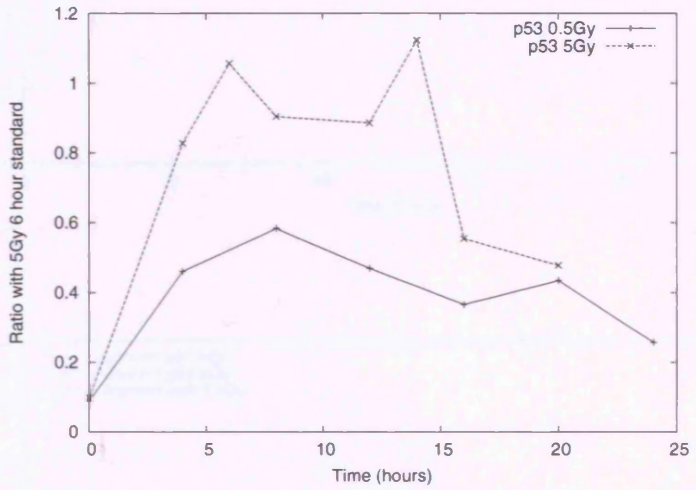
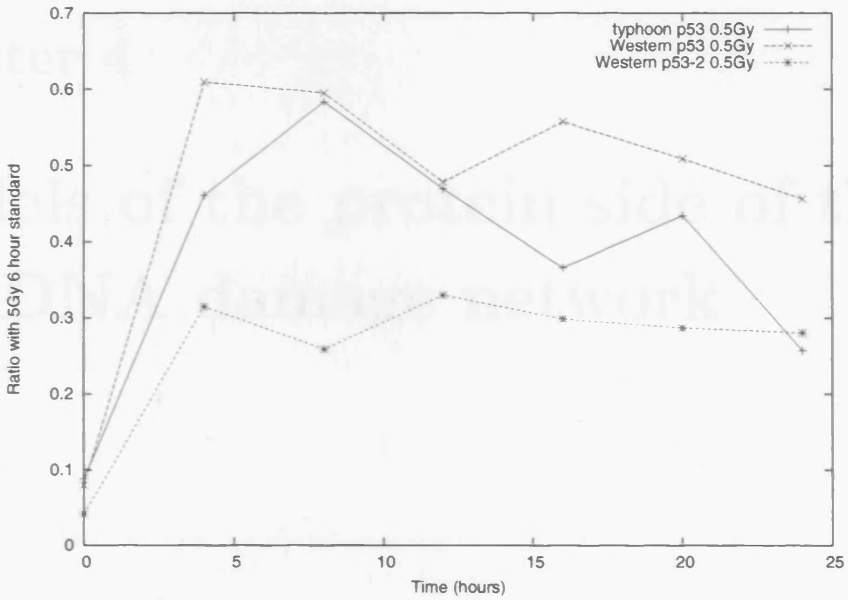


Figure 3.16: A time course plot of the response of p53 to 0.5Gy and 5Gy of radiation. The Typhoon phosphoimager was used to measure the blots.

20 hours with the possibility that there are damped oscillations. At a dose of 0.5Gy the measurements made with the Typhoon initially agree well with the first replicate of the previous Western blot method, but after 8 hours the Typhoon measurement declines more rapidly (Figure 3.17). The shape is significantly different, peaking at different places, but the time series generally remains between the two old measurements. At 5Gy of damage the shape of the Typhoon measurement is in good agreement with the second replicate of the previous method, even though the levels are generally higher for the Typhoon measurement. This gives further evidence that something erroneous occurred with the first replicate at 5Gy. These results are very good and are a significant improvement over the previous method. Unfortunately due to a lack of resources, measurements using the Typhoon for other proteins of interest were unavailable in time for this thesis.

(a)



(b)

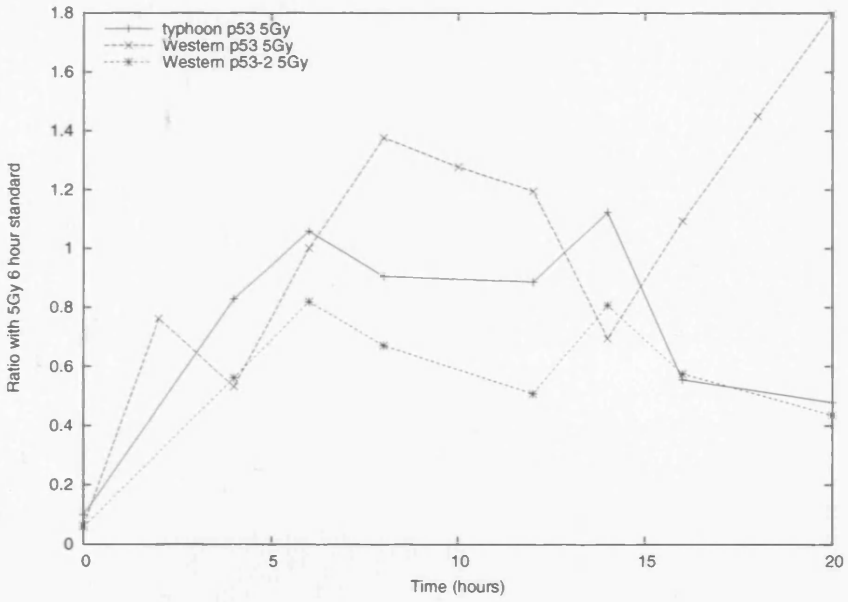


Figure 3.17: A time course plot comparing the Typhoon measurement for total p53 with two Western blot measurements using the old method at (a) 0.5Gy and (b) 5Gy of radiation.

Chapter 4

Models of the protein side of the p53 DNA damage network

4.1 Introduction

When examining data in isolation, it is often difficult to gain a full understanding of the link between the system's behaviour and its underlying components and processes. Mathematical modelling aims to overcome this by providing a simplified quantitative and predictive description of the system. A model should replicate the behaviour of the system suggested by observational data. Also it is essential that the mathematical structure of the model provides new information about the biological processes. A mathematical model is easily manipulated, which allows the full exploration of the system and the discovery of important or new mechanisms. This allows many "experiments" to be performed quickly and with nominal cost, the result of which should be predictions that can be tested in "real" experiments.

There is a wide range of modelling paradigms (de Jong, 2002; Stark *et al.*, 2003), but in this thesis non-linear ordinary differential equations will be used. This framework provides a simple but detailed description of the system, is widely used, and has a large repository of tools and techniques available. Additionally, the underlying assumptions and problems with this approach are well known. It is a reasonable compromise between simple topological structures and complex knowledge-filled stochastic or spatial approaches.

Chapter 2 outlined the main known components and interactions of the p53 gene regulatory network. In this chapter models will be proposed based on this information with the aim that, along with the data described in chapter 3, a better molecular-level description of the DNA damage response in the MOLT4 cell line will be obtained (focussing on the part of the network that regulates p53 and is disrupted by DNA damage). In particular, it would be interesting to learn whether the dynamics of p53 vary depending on the amount of DNA damage. This may help determine what role p53 plays in the decision to commit apoptosis. Also it would be interesting to discover which mechanisms are particularly important in the response, this may provide clues about which mechanisms are disrupted in cancerous cells and how these defects could be corrected.

This chapter begins with an explanation of the methodology used to construct the models followed by a description of the main models used in this thesis. Various analytically tractable "toy" models are then examined to gain a better idea of the core behaviour of the system. A full six component ODE model is proposed which would be of use if there was a greater amount of data. Finally, various problems that could potentially be answered by modelling are examined.

4.2 Setting up the mathematical models

To construct the models of the protein side of the p53 damage network the techniques used by Tyson *et al.* were generally used (Novak and Tyson, 1993; Novak *et al.*, 1998;

Tyson, 1999; Tyson and Novak, 2001). The basic premise behind these models is to take knowledge of the biochemical mechanisms of the system, simplify them and translate them into a system of non-linear differential equations.

The biochemical information introduced in section 2.2 is used as the basis for the models. The following simplifications were made: only the main known components and interactions were included, redundant pathways were cut out[1] and simple mathematical relationships are used to describe the interactions between the components. It is difficult to know how far one should simplify the system; enough complexity is needed to accurately represent the system but with too much complexity the model becomes unworkable. The approach taken here is to simplify the system as far as possible and then gradually increase the complexity.

After simplification the mechanisms are translated to non-linear differential equations based upon the law of mass action. The rate of change of a particular component will depend upon both the concentration of other system components and their interactions. There are four types of interaction that need to be described mathematically,

- Production

  There are two types of production, production that is not dependent on components within the model (basal production) and production that is dependent. Basal production is assumed to occur at a constant rate. For dependent production the rate will be assumed to be directly proportional to the concentration of the transcription factor. This does not take into account many mechanisms that would make the model more realistic. This includes the rate of production saturating because the DNA can only be transcribed so fast, co-operative transcription effects and the production of protein being a two step process.

- Degradation

  Again there are dependent degradation rates and basal degradation rates. The basal degradation rates are assumed to be directly proportional to the concentration of the component and the dependent degradation rate is assumed to be proportional to the concentration of the component and the protein that is affecting the degradation[2]. More complex effects are ignored.

- Binding

  Two components can bind together to form a third component. The rate of this process will depend on the amount of each constructing component and the rate at which they bind. A basic way to express this reaction mathematically is by a

[1] If two pathways perform the same function then they are combined. This reduces the number of components in the model and consequentially, the number of parameters that need to be estimated.

[2] There is a certain probability that a molecule will degrade and at large enough concentrations the rate of degradation can be thought of as the proportion of molecules that will degrade per unit time.

rate constant multiplied by the concentration of the two components. Here is a simple example of the rate equations where two components A and B react to form product C,

$$\frac{d[A]}{dt} = -k[A][B], \quad \frac{d[B]}{dt} = -k[A][B], \quad \frac{d[C]}{dt} = +k[A][B].$$

- Enzyme-like interactions

  Examples of enzyme-like interactions in the p53 network are phosphorylation and ubiquitination. In the Tyson *et al.* models, Michaelis-Menten equations are used to describe the enzyme action of various components. Here a simpler approach is used which requires less parameters; the rate of the interaction is assumed to be directly proportional to the amount of the "enzyme" and the amount of the target substance. For example, if substance A is converted to substance B through phosphorylation caused by substance C, then the differential equations will appear as follows,

$$\frac{d[A]}{dt} = -k[A][C], \quad \frac{d[B]}{dt} = +k[A][C], \quad \frac{d[C]}{dt} = 0.$$

There are a number of assumptions associated with ODE models. Firstly, all spatial issues have been ignored (some localisation issues will be examined later in chapter 5) and it is assumed that the system is a well mixed solution of the various components and their DNA. The second assumption is that there is enough of each substance to make differential equations realistic i.e. stochastic effects can be ignored. Finally it is assumed that the mathematical description of the biological mechanisms are accurate enough to produce the general behaviour of the system.

4.3 The model and its variants

There are six core components of the p53 gene regulatory network: p53 (active and inactive), active MDM2, active ATM, ARF and E2F1 (see chapter 2). Unfortunately, data from MOLT 4 cells were only available for four components: active p53, inactive p53, MDM2 and ATM (see chapter 3). Therefore, it is only practical to use four component models with this data. In this section a number of four component models will be introduced based on the main interactions of the network and will be used in the rest of this thesis. These models are not analytically tractable and so a number of "toy" models are examined in section 4.4 to get a better idea of the behaviour of the system. A more complete model that includes all six components is introduced in section 4.5.

The following definitions will be used through out,

$$x = [Ap53], \quad y = [MDM2], \quad z = [p53], \quad a = [ATM],$$

where $[s]$ is the concentration of component s, $MDM2$ is active MDM2, $p53$ is inactive p53 and $Ap53$ is active p53.

4.3.1 Input to the system

In all the models, active ATM drives the system. It is assumed that at time $t = 0$ the active ATM level have been "kicked" to a value away from equilibrium and this decays exponentially according to the rate constant, D_{ATM},

$$a(t) = ATM_0 e^{-D_{ATM}t},$$

where ATM_0 is the initial amount of active ATM. The initial amount of active ATM is directly associated with the initial DNA damage and is assumed to be proportional to the DNA damage. The degradation rate can be associated with the rate of DNA repair. H2AX data (section 3.3) has shown it is reasonable to assume that the initial amount of ATM is proportional to the amount of radiation the cells are exposed to and that the degradation rate will be fairly constant. The other components of the system are assumed to be at equilibrium at $t = 0$.

4.3.2 Simplifications

- The path containing CHK2 was removed. This is because ATM→CHK2→p53 duplicates the behaviour of the more direct ATM→p53 and so the effect of the CHK2 pathway can be included in the interaction between active ATM and p53.
- Once MDM2 has bound to ARF or become inactivated through phosphorylation it is removed from the system (in effect degraded). This means that only one component of MDM2 (the active form) is required.
- Active ATM is the only protein that can convert p53 into its active state and there is no "basal" rate of activation.
- If MDM2 interacts with both inactive p53 and active p53 then it degrades them at an equal rate.
- p53 forms a tetramer when activated before it can perform its function as a transcription factor. The details of this mechanism are ignored.

4.3.3 Simple four component model

A simple model (*model 1*) is proposed that includes the main interactions of the p53 network (Figure 4.1). The following interactions are modelled: through phosphorylation active ATM enables the stabilisation and hence activation of p53 (k_2); ATM phosphorylates MDM2 compromising MDM2's ability to ubiquitinate and bind p53, hence ATM

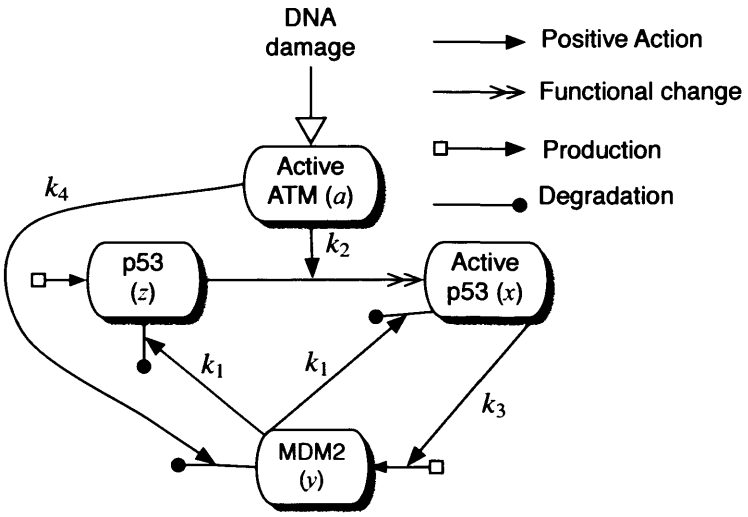


Figure 4.1: A schematic of model 1 (equation 4.1). k_i are interaction rate constants that indicate the strength of the interaction between the two components joined by the arrow.

increases the rate at which MDM2 is inactivated/degraded (k_4); active p53 transcribes MDM2 (k_3); and MDM2 encourages the degradation of both forms of p53 through ubiquitination and also prevents p53 acting as a transcription factor by binding (k_1). The model ODEs are,

		Production	Degradation	Binding/Enzyme	
$\frac{da}{dt}$	$=$		$-D_{ATM}a,$		
$\frac{dz}{dt}$	$=$	p_{p53}	$-D_{p53}z - k_1yz$	$-k_2az,$	(4.1)
$\frac{dx}{dt}$	$=$		$-D_{p53}x - k_1yx$	$+k_2az,$	
$\frac{dy}{dt}$	$=$	$p_{MDM2} + k_3x$	$-D_{MDM2}y$	$-k_4ay,$	

where,

$$
\begin{aligned}
a &= \text{Active ATM concentration} \\
z &= \text{Inactive p53 concentration} \\
x &= \text{Active p53 concentration} \\
y &= \text{MDM2 concentration} \\
k_i &= \text{Interaction rate constant } i \\
p_q &= \text{Basal production rate of } q \\
D_q &= \text{Basal degradation rate of } q
\end{aligned}
$$

This model will be used extensively in this thesis as the example model in the examination of parameter estimation techniques (chapters 6–7).

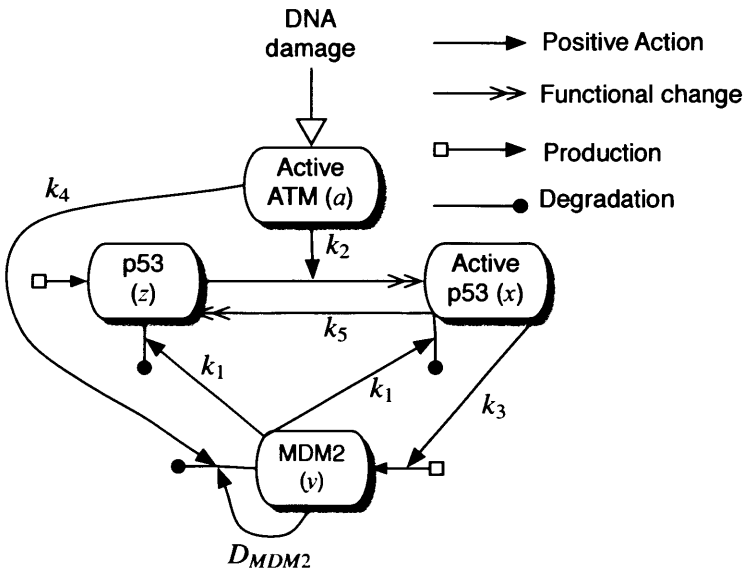


Figure 4.2: A schematic of model 2, a more complex four component model (equation 4.2).

4.3.4 More complex four component model

This model (*model 2*) is the same as model 1 (equation 4.1) but includes two additional mechanisms: the ability of active p53 to become deactivated independently (k_5) and the self-ubiquitination of MDM2 ($-D_{MDM2}y^2$) (see Figure 4.2). It is assumed in this case that the self-ubiquitination of MDM2 is the major cause of MDM2 degradation. The model equations are,

		Production	Degradation	Binding/Enzyme
$\frac{da}{dt}$	=		$-D_{ATM}a,$	
$\frac{dz}{dt}$	=	p_{p53}	$-D_{p53}z - k_1yz + k_5x$	$-k_2az,$
$\frac{dx}{dt}$	=		$-D_{p53}x - k_1yx - k_5x$	$+k_2az,$
$\frac{dy}{dt}$	=	$p_{MDM2} + k_3x$	$-D_{MDM2}y^2$	$-k_4ay.$

(4.2)

There are a number of interactions in the p53 network that have not been well established: the self-ubiquitination of MDM2, MDM2 negatively regulating inactive p53 only or both forms of p53 and active p53 becoming inactive independently[3]. It would be interesting to test, through model validation, whether these mechanisms exist and what effect they have on the dynamics. To test this a number of model variants are constructed where certain terms are removed from model 2:

1. Without MDM2 self-ubiquitination (*model 2(b)*). In this case $-D_{MDM2}y$ is used as the degradation term.

[3]Even though it is unrealistic to have no inactivation, it is possible that p53 becomes inactive at such a low rate that it is likely to be degraded before adding to the inactive pool.

2. Without ubiquitination of active p53 by MDM2 (*model 2(c)*). Here the term "$-k_1yx$" is removed from the active p53 equation.

3. Without both mechanisms (*model 2(d)*). Term "$-k_1yx$" is removed and $-D_{MDM2}y$ is used as the sole MDM2 degradation term.

To evaluate whether the independent deactivation mechanism of active p53 produces a significant better fit to the data, model 2 and model 2(d) could be compared; this might also give an indication of the time spent in an active state. An attempt to examine these mechanisms though model validation is performed in section 7.8.

4.4 Analytically tractable models and stability analysis

To get an idea about the general properties and behaviour of the protein side of the p53 network, various analytically tractable models are examined. Initially the simplest possible models will be constructed and gradually made more complex. In this analysis ATM (a) is treated as being absent or as a parameter by using the quasi-steady state assumption.

4.4.1 Simple models

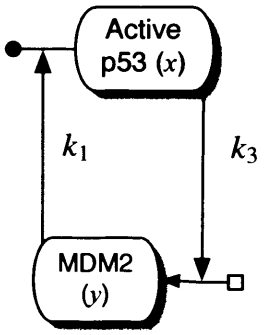


Figure 4.3: Schematic of the p53/MDM2 negative feedback loop

To begin a simple model of the core of the p53 network is proposed, which is a basic negative feedback loop with a non-linearity in the p53 degradation term (see Figure 4.3). ATM is initially ignored and there is no basal production or degradation. The model is described in ODEs as follows,

$$\begin{aligned} \dot{x} &= -k_1yx, \\ \dot{y} &= k_3x. \end{aligned}$$

Linear stability analysis is performed by examining the properties of the Jacobian matrix (Strogatz, 2000). The results give a "line" of fixed points with $x^* = 0$ and y^*

undefined. This is unrealistic biologically and suggests the model is too simple. ATM is now introduced but is assumed to be constant and hence a parameter (quasi-steady state approximation). ATM has a positive influence on the amount of active p53 and a negative effect on the amount of MDM2,

$$\begin{aligned} \dot{x} &= -k_1 y x + \alpha_1 a, \\ \dot{y} &= k_3 x - \alpha_2 y a. \end{aligned}$$

The equilibrium fixed points are,

$$\begin{aligned} x^* &= \pm\sqrt{\frac{\alpha_1\alpha_2}{k_1 k_3}}a, \\ y^* &= \pm\sqrt{\frac{\alpha_1 k_3}{k_1\alpha_2}}. \end{aligned}$$

As x and y are only defined when positive there is one fixed point. Interestingly, in this situation the equilibrium level of MDM2 does not depend on the amount of ATM. The fixed point is stable and it can be classified as either a spiral (damped harmonic oscillations) or a node (steadily approaching equilibrium) depending on the amount of ATM. If,

$$(3-\sqrt{8})\sqrt{\frac{k_1 k_3 \alpha_1}{\alpha_2^3}} < a < (3+\sqrt{8})\sqrt{\frac{k_1 k_3 \alpha_1}{\alpha_2^3}},$$

then the fixed point is a spiral. Even at this simplistic level possible mechanisms to explain data from experiments are apparent. When the cell is not stressed and ATM levels are low, the amount of active p53 is low. When DNA damage occurs, ATM levels are high and the level of active p53 are proportionally higher. The level of damage depends on what dynamics are observed; when the ATM level is very high or very low the equilibrium is approached steadily but at intermediary levels there are damped oscillations. If for the cell type used in Bar-Or *et al.* (2000) (MCF-7) high levels of damage caused only intermediate levels of ATM, this model might explain their results, where oscillations only occurred at high levels. For the data gathered for this project (see chapter 3), oscillations seem to occur only at low levels of damage. This could be explained by the model if for MOLT4 cells a low amount of damage caused an intermediary level of ATM and a high amount of damage caused a high level of ATM. A model prediction is that MOLT4 cells are more sensitive to DNA damage than MCF-7 cells. The behaviour of MDM2 is not well explained by this model.

4.4.2 Adding basal production and degradation

To examine a different aspect of the p53 system a basal rate of p53 production and a basal degradation rate for MDM2 were added but ATM was removed (Figure 4.4). The

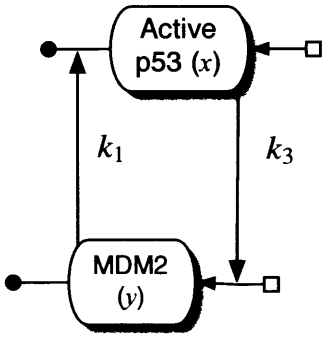


Figure 4.4: Schematic of a simple model with added basal rates

corresponding ODEs are as follows,

$$\begin{aligned} \dot{x} &= p_{p53} - k_1 y x, \\ \dot{y} &= k_3 x - D_{MDM2} y. \end{aligned}$$

There is one stable positive fixed point at,

$$x^* = \sqrt{\frac{p_{p53} D_{MDM2}}{k_1 k_3}}, \quad y^* = \sqrt{\frac{p_{p53} k_3}{k_1 D_{MDM2}}}.$$

The effect of DNA damage can be simulated by an increase in p_{p53} and D_{MDM2}, and a decrease in k_1. This will increase the equilibrium amount of p53 and increase or decrease MDM2 depending on the relative change of these parameters. A variety of behaviour can be observed depending on the parameter values, one interesting inequality is that if,

$$\frac{(17 - 12\sqrt{2}) D_{MDM2}^3}{k_1 k_3} < p_{p53} < \frac{(17 + 12\sqrt{2}) D_{MDM2}^3}{k_1 k_3},$$

the fixed point is a spiral, otherwise it is a node. DNA damage appears to have the same effect as in the previous model with damped oscillations only occurring at intermediate levels of damage. As this model does not contain ATM as a component it is impossible to know what the dynamics would be like after damage. In particular, it cannot be known whether the oscillations proposed are similar to those observed in experiments and if p53 can increase after damage at the rate observed.

4.4.3 Active and inactive p53 without MDM2

To examine the effects of p53 activation, MDM2 is removed as an intermediary step and two states of p53 are introduced. The effect of MDM2 is represented by active p53 having a negative effect on the amount of p53 (see Figure 4.5). There is some debate about whether MDM2 interacts differently or not at all with active p53 (Stommel *et al.*, 1999; Gu *et al.*, 2001). In this model it is assumed that MDM2 does not interact with active

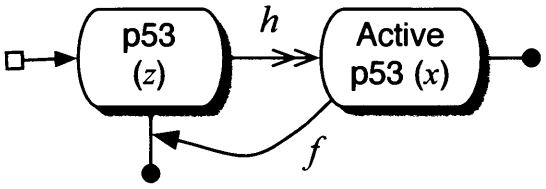


Figure 4.5: Schematic of model that includes p53 and active p53 but not MDM2.

p53. if it did interact then the effect would be introduced by active p53 autoregulation. The model equations are,

$$\begin{aligned} \dot{z} &= p_z - D_{p53}z - hz - fzx, \\ \dot{x} &= -D_{p53}x + hz. \end{aligned}$$

There is one real positive fixed point which is stable. There is no possibility of oscillations in this system. suggesting that an intermediary is required to get this kind of behaviour.

4.4.4 MDM2 autoregulation

There is some evidence that MDM2 down-regulates itself through self ubiquitination (Fang *et al.*, 2000). this could have profound effects on the system and is likely to increase the recovery rate when MDM2 levels are high after DNA damage. A simple model

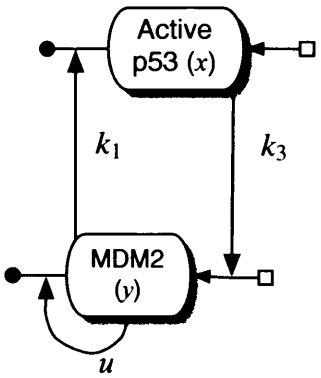


Figure 4.6: Schematic of a simple model with MDM2 autoregulation

(Figure 4.6) that represents this mechanism was constructed with the following ODEs,

$$\begin{aligned} \dot{x} &= p_{p53} - D_{p53}x - k_1yx, \\ \dot{y} &= p_{MDM2} - D_{MDM2}y + k_3x - uy^2 \end{aligned}$$

This produces complicated results. There is one stable real positive fixed point that is approached through damped harmonic oscillations in certain situations.

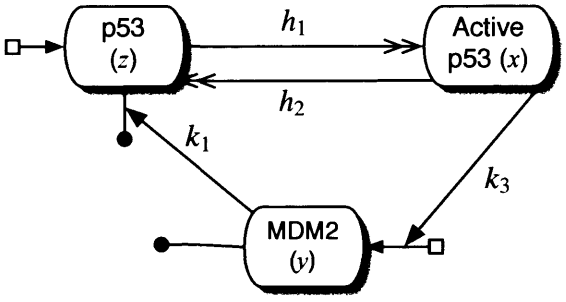


Figure 4.7: Schematic of a simple three component model with p53 deactivation.

4.4.5 Deactivation of p53

In this model, there are three components, p53, active p53 and (active) MDM2. Basal rates are only included if there is no other component that causes that mechanism. The focus of this model is that p53 becomes inactive at a certain rate (Figure 4.7). This is a more realistic scenario then previous models where once active it stays active. Again, the assumption is used that MDM2 only degrades inactive p53. The model equations are,

$$\begin{aligned} \dot{z} &= p_{p53} + h_2 x - h_1 z - k_1 z y, \\ \dot{x} &= -h_2 x + h_1 z, \\ \dot{y} &= k_3 x - D_{MDM2} y. \end{aligned}$$

There is only one positive real equilibrium point and it is stable,

$$z^* = \sqrt{\frac{D_{MDM2} h_2 p_{p53}}{h_1 k_1 k_3}}, \quad x^* = \sqrt{\frac{D_{MDM2} h_1 p_{p53}}{h_2 k_1 k_3}}, \quad y^* = \sqrt{\frac{h_1 k_3 p_{p53}}{D_{MDM2} h_2 k_1}}$$

DNA damage has the effect of increasing the rate of p53 activation (h_1), decreasing the rate of ubiquitination by MDM2 (k_1) and increasing the rate of inhibition of MDM2 (D_{MDM2}). This would increase the equilibrium concentration of active p53 but the concentration level of other components would depend on the ratio of the altered parameters. Unfortunately the stability analysis is too complex to examine in detail but numerical simulations reveal that damped oscillation occur at certain parameter values.

4.4.6 Summary

It was only possible to analyse simple models and even then the answers were sometimes too complex to discern a clear relationship. Despite this, some interesting results have been obtained and the models replicate much of the observed behaviour. The principal result is that it is possible for the core of the network to produce oscillations, which have been observed experimentally. The presence of MDM2 is a requirement and it appears that oscillations only occur at intermediate levels of damage. From this it was

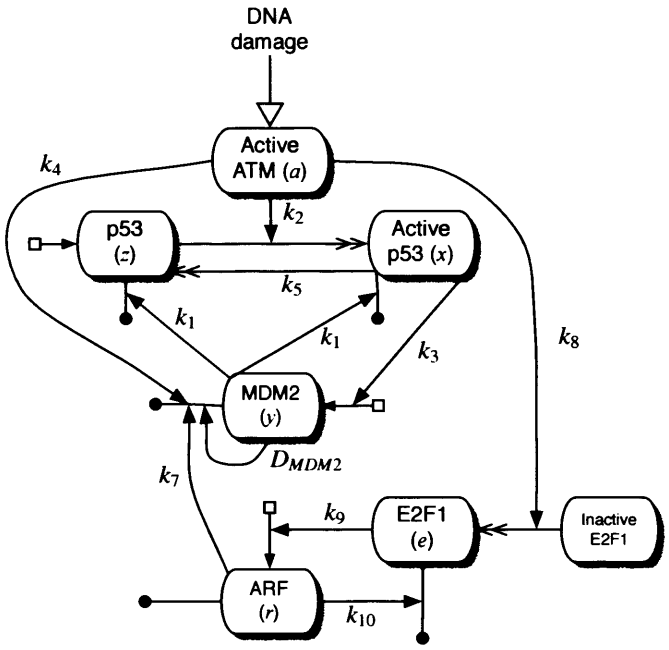


Figure 4.8: Schematic of the full six component model

possible to predict that MOLT4 cells are more sensitive to DNA damage than MCF-7 cells. Additionally, in all cases DNA damage caused an increase in the amount of active p53, as is known to occur. Finally, there seems to be only a single stable fixed point in the core of the p53 system. There have been no observations of bistable behaviour in the p53 system and biologically there needs to be at least one stable fixed point.

These models are too simple to gain a complete insight into the DNA damage response. This is mainly because the input to the system is not correctly defined and so it is impossible to capture what the dynamics after damage would be like. In particular it is unclear whether it is possible for active p53 to rise as quickly as is observed experimentally, whether MDM2 has the correct dynamics, and whether the oscillations are of the same period and amplitude.

4.5 The full model

The main models used in this thesis are restricted to four components due to the data that was available. In this section a model is proposed that uses all the core components by adding ARF and E2F1 (Figure 4.8). ATM activates E2F1 (k_8) allowing it to transcribe ARF (k_9). ARF in turn binds to E2F1 (k_{10}), thus producing another negative feedback loop. ARF also binds to MDM2 ($k7$) making it inactive in the system. It was assumed that once ARF binds with E2F1 or MDM2 both proteins are removed from the system. It was also assumed that there is an infinite pool of inactive E2F1, so that at no time is there a lack of inactive E2F1 to convert to active E2F1. This removed the need to have

inactive E2F1 as a component of the model. A simple version of the model is described as,

$$\begin{array}{lcll} & & \text{Production} & \text{Degradation} & \text{Binding/Enzyme} \\ \frac{da}{dt} & = & & -D_{ATM}a, & \\ \frac{dz}{dt} & = & p_{p53} & -k_1yz & -k_2az, \\ \frac{dx}{dt} & = & & -k_1yx & +k_2az, \\ \frac{dy}{dt} & = & p_{MDM2}+k_3x & -D_{MDM2}y & -k_4ay-k_7ry, \\ \frac{dr}{dt} & = & p_{ARF}+k_9e & -D_{ARF}r & -k_7ry-k_{10}re, \\ \frac{de}{dt} & = & k_8a & -D_{E2F1}e & -k_{10}re, \end{array} \tag{4.3}$$

where r is the concentration of ARF and e is the concentration of active E2F1. A more complex version of this model is described as follows (including MDM2 self ubiquitination and independent active p53 inactivation),

$$\begin{array}{lcll} & & \text{Production} & \text{Degradation} & \text{Binding/Enzyme} \\ \frac{da}{dt} & = & & -D_{ATM}a, & \\ \frac{dz}{dt} & = & p_{p53} & -D_{p53}z-k_1yz+k_5x & -k_2az, \\ \frac{dx}{dt} & = & & -D_{p53}x-k_1yx-k_5x & +k_2az, \\ \frac{dy}{dt} & = & p_{MDM2}+k_3x & -D_{MDM2}y^2 & -k_4ay-k_7ry, \\ \frac{dr}{dt} & = & p_{ARF}+k_9e & -D_{ARF}r & -k_7ry-k_{10}re, \\ \frac{de}{dt} & = & k_8a & -D_{E2F1}e & -k_{10}re. \end{array}$$

ARF is known to be an important regulator of MDM2, so it is likely that the mechanisms introduced in this model will be important. Of particular interest would be the balance between the two feedback loops before and after damage. The simple model (equation 4.3) along with some variants were the focus of some previous work (Brewer, 2002). It was found that depending on the parameters there was a large range of behaviour (including damped oscillations) but there was never sustained oscillations. The general effect of an increase of ATM was a decrease in the concentration of inactive p53 and an increase in the concentration of active p53. Only approximately 3% of parameter sets tested caused oscillations but it was found that some parameters, in particular k_4, k_1, k_7, D_{MDM2}, p_{p53} and p_{MDM2}, play an important role in determining whether oscillations were produced. It was also determined that there was no sharp switch to oscillatory behaviour as ATM was increased from a low level.

4.6 Interesting problems that arise from p53 experiments

There is some evidence that p53 and MDM2 oscillate after DNA damage is experienced by the cell (see sections 2.4.1, 2.4.3 and 3.5.2). It is not yet clear whether the oscillations are sustained or damped, or whether they have any biological significance. At the population level the oscillations seem to be damped but initial evidence suggests that at

the single cell level they are sustained (Lahav *et al.*, 2004). In mathematical terms sustained oscillations means that there must be a stable limit cycle, which normally results from a Hopf bifurcation. For sustained oscillations to occur there needs to be a negative feedback loop, but this is not sufficient (Fall *et al.*, 2002); if there are two components then Bendixson's criterion requires that one component is autocatalytic (there is a positive feedback loop). If there are more than three components in the negative loop then sustained oscillations are possible with only negative feedback. Oscillations between p53 and MDM2 do not occur in unstressed cells which suggests DNA damage causes this dynamic. Both Ciliberto *et al.* (2005) and Ma *et al.* (2005) have shown that it is possible for the p53 network to produce sustained oscillations (see section 2.4.4). Monk also showed that transcriptional delay could cause oscillations (Monk, 2003a). The data though is not strong enough to dismiss the notion that damped oscillations occur between p53 and MDM2 within the single cell. Damped oscillations were found to occur in a number of the "toy" models proposed and in the full model. These oscillations are interesting as they may serve a biologically purpose. Do the oscillations play a role in the decision to commit apoptosis? or is it a delaying technique to allow the cell every opportunity to repair itself?

The p53 network controls whether a cell commits apoptosis or not, this means that at some stage there is probably a switch. There are a number of places in the network where this switch could occur,

1. At the damage signal level

   In this situation there is no switch as such but the damage signal is strong enough and lasts long enough so that the apoptosis machinery reaches such a concentration that apoptosis occurs. Dynamically the system does not switch to the basin of attraction of another equilibrium point.

2. Between the damage signal and p53

   The system would need to be bistable; there would be two states, one, the equilibrium state, where all components are relatively low and an active state where the equilibrium concentration of active p53 at least would be high. The damage signal, if it is strong enough, would displace the system into the active state and the level of active p53 would remain high even after the damage signal ceases (probably hysteresis).

3. Downstream from p53

   The switch could occur in one or a number of the sub-systems that p53 affects. In this situation once the damage signal has stopped p53 drops back to pre-damage levels, but a sub-system p53 affects remains at a high level of activity, leading to apoptosis.

Bistability and more generally multistationarity relies on positive feedback existing within the system (Thomas and Kaufman, 2001). It does not appear that there is any positive feedback in the core of the p53 network and so none of the models that are developed here are sufficient to produce switch like behaviour. This suggests that the decision between whether to commit to apoptosis or not is not based in this part of the network. Ciliberto *et al.* (2005) have found a positive feedback loop but it remains unclear whether it actually impacts the core of the system .

Chapter 5

Analysis of a localisation model of the DNA damage p53 gene regulatory network

5.1 Introduction

Over the last few years it has become increasingly clear that a key mechanism in the regulation of the p53 network response to stress is the control of the location of the network's principal components, in particular p53 and MDM2 (O'Brate and Giannakakou, 2003; Liang and Clarke, 2001; Michael and Oren, 2003). It has been shown that in a subset of tumour cells such as breast cancers, colon cancers and neuroblastoma, wild-type p53 is abnormally confined to the cytoplasm (Liang and Clarke, 2001). These tumours are found to be less responsive to cancer treatments such as radiotherapy and chemotherapy. This is because when p53 is confined to the cytoplasm it does not have access to DNA so cannot transcribe pro-apoptotic genes, preventing one of the major routes to apoptosis. The models developed so far were not designed to address these issues. Therefore, in this chapter a model that includes some of these localisation mechanisms will be examined to determine how these mechanisms improve the response of the p53 network.

There are two essential elements of a protein that enable it to be actively transported from the nucleus to the cytoplasm and *vice versa*: the nuclear localisation signal (NLS) which enables nuclear import, and the nuclear export signal (NES) which enables nuclear export. Protein p53 has three NLSs (Shaulsky *et al.*, 1990), only one of which is found to be strongly active, and two NESs (Stommel *et al.*, 1999; Zhang and Xiong, 2001). One of these NESs has been found to be necessary and sufficient to direct p53 nuclear export (Stommel *et al.*, 1999). When the cell is not under stress p53 is a short-lived protein due to rapid degradation and most of p53 is found in the cytoplasm (Li *et al.*, 2003; Liang and Clarke, 2001). In this situation the net effect is that p53 is exported, keeping nuclear p53 to a minimum, protecting the cell from any apoptotic effects. The exception to this is at the G1/S phase transition where p53 does enter the nucleus (Hayon and Haupt, 2002). When the cell does experience stress the situation is reversed with the majority of p53 being located in the nucleus (O'Brate and Giannakakou, 2003). This could be caused by either an increase in nuclear input or a decrease in nuclear export but the evidence points to it being a decrease in export. For p53 to become an active transcription factor it must undergo a conformational change and form a tetramer (Vogelstein *et al.*, 2000). The rate at which p53 is activated increases when the cell experiences stress and it has been shown that active p53 does not get exported from the nucleus. The export is prevented in two ways, firstly, when p53 forms a tetramer the NES is covered thus blocking nuclear export (Stommel *et al.*, 1999). Secondly, when p53 is phosphorylated on serine 15 or 20 this not only causes a conformational change but prevents nuclear export (O'Brate and Giannakakou, 2003) (probably by affecting the accessibility of the NES (Liang and Clarke, 2001)). Active ATM, CHK1, CHK2 and JNK all phosphorylate p53 on these sites (Appella and Anderson, 2001) and their concentrations are also increased by different routes of stress.

MDM2 is the most significant negative regulator of p53. MDM2 marks p53 for degra-

dation by ubiquitinating it and also binds to the transcriptional activation site of p53 preventing it performing its function (Vogelstein *et al.*, 2000). MDM2 also regulates p53 by increasing the nuclear export of p53, removing it from its area of function (Liang and Clarke, 2001). It is now generally accepted that this occurs by MDM2 binding and ubiquitinating p53 which exposes the NES, allowing p53 to be exported and then degraded (Stommel *et al.*, 1999; Li *et al.*, 2003). It has been shown that MDM2 ubiquitin ligase activity is required for p53 nuclear export (Kawai *et al.*, 2003) and that MDM2-dependent p53 nuclear export requires an intact NES in p53 but not in MDM2 (Boyd *et al.*, 2000; Geyer *et al.*, 2000). The other possible export mechanism is that MDM2 binds to p53 and carries it out of the nucleus (O'Brate and Giannakakou, 2003).

An interesting and important set of results was obtained by Li *et al.* (2003). They found that there were two distinct types of behaviour between p53 and MDM2 depending on the quantity of MDM2. At low amounts of MDM2, p53 is mono-ubiquitinated (this can be at multiple sites) and nuclear export occurs, whereas at high levels of MDM2, p53 is poly-ubiquitinated and there is rapid degradation. This behaviour was confirmed both *in vitro* and *in vivo*. When there was no MDM2, p53 was mainly confined to the nucleus but when there were low amounts of MDM2 the p53 was mainly confined to the cytoplasm. When there were high levels of MDM2 the amount of p53 was undetectable but if the cells were treated with proteasome inhibitors it was found that p53 was mainly confined to the nucleus. This suggests that the degradation is occurring in the nucleus, but there is still considerable debate about where p53 is principally degraded. It has been shown that there can be between 1 and 6 ubiquitin tags in a p53 ubiquitination chain (Lai *et al.*, 2001). So p53 is dealt with in two separate ways depending on the number of ubiquitin tags it has. Both mechanisms cause p53 to become inactive but when p53 is mono-ubiquitinated it is inactive only temporarily. The mechanisms probably have two functional uses. It could be because the mono-ubiquitination mechanism has a low energy cost compared with the poly-ubiquitination. When a cell has just recovered from cell stress (and repaired any damage) levels of both MDM2 and p53 will be high and it is a matter of urgency to remove the excess p53 before apoptosis is initiated, therefore destroying the p53 using the relatively energy expensive poly-ubiquitination is reasonable. In a normal situation wasting energy through poly-ubiquitination and destroying p53 is not justified and so p53 is just removed from the nucleus. Another possibility is that these mechanisms provide an improved response to stress. The active nuclear export means that there is a "reserve" of p53 in the cytoplasm, thus when the cell experiences stress the nuclear export is stopped and the p53 can flood into the nucleus. This will cause a faster response. It is principally the possibility of these mechanisms that will be examined in the models. It has also been suggested that MDM2 promotes its own decay effectively through poly-ubiquitination (Shmueli and Oren, 2004)[1].

[1]This will not considered in the models, but it would introduce a non-linearity in MDM2 decay meaning that when MDM2 levels were high it would more rapidly degrade than when levels were low.

ARF is another major component of the p53 gene regulatory network and functions as a negative regulator of MDM2. Not only does ARF bind and inactivate MDM2 but also moves MDM2 into the nucleoli (Tao and Levine, 1999; Weber *et al.*, 1999). This physically separates MDM2 from p53, allowing p53 to remain in the nucleus. There is also some evidence that, depending on cell type, the pro-apoptotic protein BCL2, which is a target of p53, can inhibit the nuclear import of p53 (Ryan *et al.*, 1994). Finally, PI3K and Akt have been reported to activate the nuclear import of MDM2 by phosphorylation and hence they are negative regulators of p53 (O'Brate and Giannakakou, 2003).

The localisation of p53 and other members of the p53 gene regulatory network plays a key role in the functioning of the network. Of particular interest are the two fates of p53 that depend on the amount of ubiquitination by MDM2. In this chapter models will be proposed based on the above information and these will be examined to gain insight into the localisation mechanism and its affect on the functioning of the network. Three increasing complex models will be analysed. Firstly, a simple chain model that includes inactive p53 in its various forms but does not include active p53 or MDM2. Secondly, the pulse model, a model that builds on the chain model and includes active p53 and MDM2. It does not explicitly include ATM/DNA damage and will be analysed by varying parameter values. Finally the complete model that includes ATM as a variable. For each model a corresponding null model will be proposed that will be same as the model of that section but will not have the nuclear export mechanism; this will allow the determination of what effect the export mechanism has. Throughout this chapter the results are basically qualitative because the parameter values can only be estimated approximately.

5.2 Simplifications and assumptions

The cell will be modelled using ODEs in two compartments, the nucleus and the cytoplasm. Each compartment is regarded as a well mixed solution with a uniform distribution of components. This simplifies the system by removing the need to implement the complex physics of import and export and solving partial differential equations (PDEs).

Three components of the p53 model will be considered: active ATM, p53 and MDM2. p53 and MDM2 are the core components of the network and active ATM is the input signal into the network for DNA damage. p53 will be considered in a number of different forms; inactive p53, active p53 and various tagged forms. Active p53 is defined as p53 that can function as a transcription factor, it is assumed active p53 is in tetramer form and becomes activated in one step (the intermediary steps of phosphorylation, conformational change and binding to other p53 proteins are not considered). This p53 ubiquitination system is simplified by assuming there are effectively two p53 ubiquitinated states:

1. p53-tag. In this state p53 has been mono-ubiquitinated at least once and is tagged to be exported from the nucleus but not degraded.

2. p53-tag-tag. In this state p53 has a poly-ubiquitinated chain so long that it has been signalled to be destroyed.

Additionally, it is assumed that both of these states of p53 do not exist in the cytoplasm i.e. when p53-tag gets exported from the nucleus it loses its ubiquitin tag and p53-tag-tag is quickly degraded (in cytoplasm or nucleus) so it has no impact. Another simplification is that that p53 can only be degraded through this route, there is no basal (outside the system being studied) degradation rate.

MDM2 will be regarded as functionally active and only localised in the nucleus; all localisation effects concerning MDM2 will be ignored. It is assumed that MDM2 is only produced through the action of active p53, there is no basal production rate. MDM2 is the only component that ubiquitinates p53, moving p53 to p53-tag and p53-tag to p53-tag-tag. The rate of these transitions are considered equal. Also it is assumed that the ubiquitin tags cannot be removed within the nucleus. Active ATM is used as the input signal to the system and it is assumed that the ATM level is proportional to the DNA damage.

Diffusion of p53 across the nucleus boundary is ignored and instead it is assumed that p53 only moves in and out of the nucleus by active transport. Diffusion can occur across the nucleus boundary if the size of the protein is less than approximately 50 kDa (Talcott and Moore, 1999). p53's size is 53 kDa, so diffusion is unlikely to have a major impact. Vousden agrees that active transport is likely to be more important (Vousden and Woude, 2000). p53 export through ubiquitination is considered to be the only way p53 is exported from the nucleus, there is no basal rate of export.

As in previous models all the interactions will be described in the simplest possible way (see chapter 4). In addition various intermediary steps have been removed and duplicate pathways combined. For example, active ATM's affect on both p53 and MDM2 combines not only the ATM protein's affect but the effect of those intermediaries that ATM activates.

5.3 Simple chain model

5.3.1 Setup

In this section, a simple chain model of the localisation mechanism will be constructed and examined (the *chain model*). This will be the first step in building more complex models but will allow some initial analysis. p53 is constructed in the cytoplasm and then transported into the nucleus at a rate of ρ, where it is ubiquitinated at a rate of α into its first and second ubiquitinated state (Figure 5.1). p53-tag is actively exported out of the nucleus at a constant rate, k. When there is DNA damage, active ATM levels rise, preventing MDM2 ubiquitinating p53; this has the effect of reducing the rate α. This

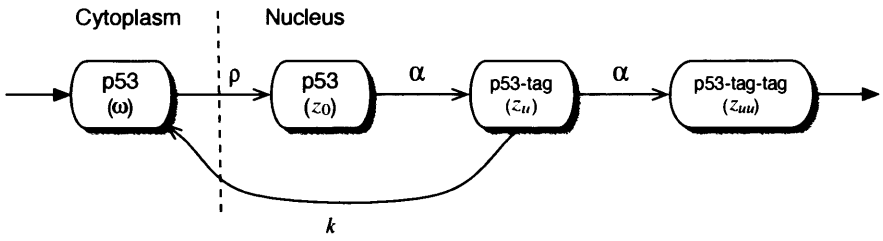


Figure 5.1: A schematic of a simple p53 localisation chain model.

model's associated ordinary differential equations are as follows,

$$\begin{aligned}
\frac{d\omega}{dt} &= p_z + kz_u - \rho\omega, \\
\frac{dz_0}{dt} &= \rho\omega - \alpha z_0, \\
\frac{dz_u}{dt} &= \alpha z_0 - kz_u - \alpha z_u, \\
\frac{dz_{uu}}{dt} &= \alpha z_u - \beta z_{uu},
\end{aligned} \tag{5.1}$$

where ω is the concentration of cytoplasmic p53, z_0 is the concentration of nuclear p53 with no ubiquitin tags, z_u is the concentration of nuclear p53-tag and z_{uu} is the concentration of p53-tag-tag. p_z is the production rate of p53 and β is the degradation rate of p53 when it is in the double tagged state. The simplest type of rate relation has been used with the assumption that there is no saturation or other non-linear effects. In this case the equations are linear. For comparison, a null model is constructed that does not have the localisation mechanisms (*null model 1*) and is described as follows,

$$\begin{aligned}
\frac{d\omega}{dt} &= p_z - \rho\omega, \\
\frac{dz_0}{dt} &= \rho\omega - \alpha z_0.
\end{aligned} \tag{5.2}$$

The only important variable is nuclear p53, but cytoplasmic p53 needs to be included so that the two models can be compared. Once nuclear p53 has been tagged it is on a non-reversible pathway to degradation so holds little interest.

5.3.2 Results and discussion

The equilibrium conditions for the chain model (equation 5.1) are,

$$\begin{aligned}
\omega^* &= \frac{p_z}{\rho}\left(1 + \frac{k}{\alpha}\right), \\
z_0^* &= \frac{p_z}{\alpha}\left(1 + \frac{k}{\alpha}\right), \\
z_u^* &= \frac{p_z}{\alpha}, \\
z_{uu}^* &= \frac{p_z}{\beta}.
\end{aligned}$$

For null model 1 (equation 5.2) the equilibrium conditions are,

$$\begin{aligned} \omega^* &= \frac{p_z}{\rho}, \\ z_0^* &= \frac{p_z}{\alpha}, \end{aligned}$$

which are the same as the equilibrium conditions for the chain model if $k = 0$. Both of these sets of equilibria are stable as long as the parameters are positive (the trace of the Jacobian is negative and the determinant is positive). It is clear that $z_0^* > z_u^*$ always and $\omega^* > z_0^* > z_u^* > z_{uu}^*$ if $\alpha > \rho$ and $\beta > \alpha$. A normal cell (not under stress) has the majority of p53 in the cytoplasm, which suggests that α is greater than ρ.

There is a pool of additional p53 in the cytoplasm because active nuclear export occurs. The size of this pool is,

$$\frac{p_z k}{\alpha \rho}.$$

The rate of nuclear export, k, increases the size of the pool as it is increased and is the main differentiator of the pool size from the other components. k drives the additional imbalance in the chain so it has a major effect. DNA damage has the effect of decreasing α. As α approaches zero the size of the pool increases approaching infinity. This disagrees with biological experiments which indicate that when there is stress, the majority of p53 is in the nucleus. This probably occurs because active p53 is not considered in this model and when there is stress the majority of p53 is active.

The amount of p53 in the nucleus is synonymous in this model with the amount of active p53. As α decreases z_0^* increases as is expected after damage. The difference between the equilibrium amount of nuclear p53 in the chain model and null model 1 is kp_z/α^2 (Figure 5.2). If either k is small or α is large the difference is small. This suggests that the export mechanism makes a significance difference in the levels of active p53 when the cell feels stress. The greater the rate of export the larger this difference.

When the cell is under stress the system is likely to be far away from equilibrium so these equilibrium equations do not apply, but they do provide the overall direction the system will move in. For example, if α is suddenly reduced, then z_0^* will increase and so at that point z_0 will be generally increasing. Suppose that the system is in equilibrium with $\alpha = \alpha_1$ and suddenly the rate of ubiquitination changes to $\alpha = \alpha_2$. In this situation the instantaneous rate of change of nuclear p53 in the chain model will be,

$$\frac{dz_0}{dt} = p_z \left(1 + \frac{k}{\alpha_1}\right)\left(1 - \frac{\alpha_2}{\alpha_1}\right),$$

whereas for the null model it will be,

$$\frac{dz_0}{dt} = p_z \left(1 - \frac{\alpha_2}{\alpha_1}\right).$$

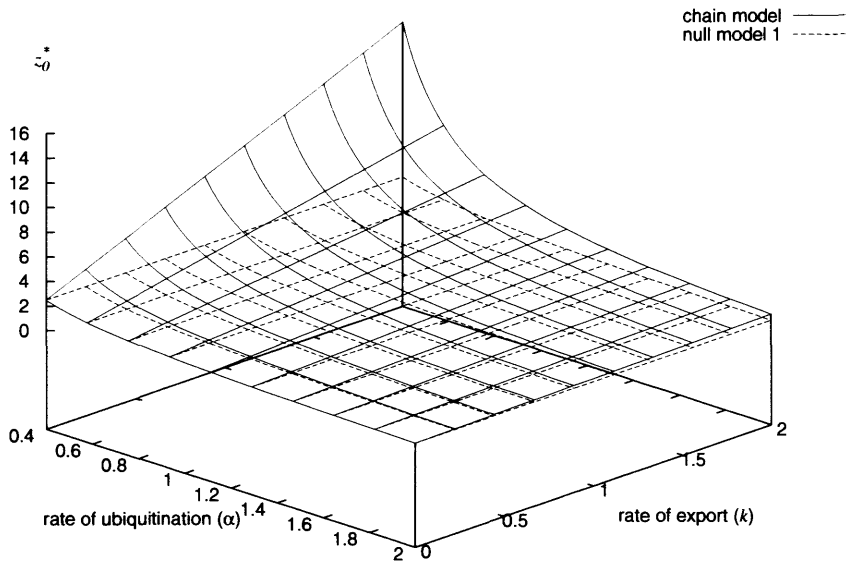


Figure 5.2: A plot showing the variation of the equilibrium concentration of z_0 with parameters k and α for the chain model (equation 5.1) and null model 1 (equation 5.2). $p_z = 1$.

The difference between the two rates is,

$$\frac{p_z k}{\alpha_1}\left(1 - \frac{\alpha_2}{\alpha_1}\right).$$

Therefore, if $\alpha_2 < \alpha_1$ (as would occur after DNA damage), the active export mechanism increases the rate at which the system can respond, providing a possible reason why the cell has this function. The larger the difference between the αs the bigger the improvement in the rates so the mechanism is particularly effective when there is a large amount of damage. The difference in rates is proportional to k suggesting that the greater the nuclear export, the larger the possible response speed is. The difference in rates is also proportional to p_z; p_z is known to be high, suggesting a real improvement caused by the active export mechanism. The difference is inversely proportional to α_1, so when α_1 is large this dampens the effect. This would be the situation prior to DNA damage (from a design point of view this is counter-intuitive). If $\alpha_2 > \alpha_1$, the difference is negative, but now both rates are negative so the magnitude of change will still be greater for the chain model. This suggests that the export mechanism may help the system recover faster from damage.

5.4 Model with active p53 and MDM2

5.4.1 Setup

Now active p53 and MDM2 are added as variables. ATM is not explicitly included as a variable, but its affect will be simulated by changing parameter values. This has the

effect of assuming that DNA damage comes in a square pulse i.e. its either on or off. This model is labelled the *square pulse model* (Figure 5.3).

When p53 is in the nucleus, it is converted to the active form of p53 at a rate of α_2 and converted back to an inactive form at a rate D_x. The activation rate is proportional to z_0^4 which is the relationship suggested for tetramerisation in a well-mixed solution of molecules. Once in its active form, p53 transcribes MDM2 at a rate of k_2. MDM2 degrades at a rate of $\alpha_3 + D_y$, where D_y is the basal rate. MDM2 ubiquitinates p53 into its two tagged states at a rate α_1. In this model, when there is DNA damage, active ATM levels rise, preventing MDM2 ubiquitinating p53, converting p53 into its active form and removing MDM2 from its function; this has the effect of reducing the rate α_1 and increasing the rates α_2 and α_3. This model's associated ordinary differential equations are as follows,

$$\begin{aligned}
\frac{d\omega}{dt} &= p_z + kz_u - \rho\omega, \\
\frac{dz_0}{dt} &= \rho\omega - \alpha_1 z_0 y - 4\alpha_2 z_0^4 + 4D_x x, \\
\frac{dz_u}{dt} &= \alpha_1 z_0 y - kz_u - \alpha_1 z_u y, \\
\frac{dz_{uu}}{dt} &= \alpha_1 z_u y - \beta z_{uu}, \\
\frac{dx}{dt} &= \alpha_2 z_0^4 - D_x x, \\
\frac{dy}{dt} &= k_2 x - (D_y + \alpha_3) y,
\end{aligned} \tag{5.3}$$

where x is the concentration of active p53 and y is the concentration of MDM2. z_{uu} plays no active role in the system and so will be ignored in the analysis below.

The corresponding null model (*null model 2*), which contains no export mechanism, is constructed as follows,

$$\begin{aligned}
\frac{d\omega}{dt} &= p_z - \rho\omega, \\
\frac{dz_0}{dt} &= \rho\omega - \alpha_1 z_0 y - 4\alpha_2 z_0^4 + 4D_x x, \\
\frac{dx}{dt} &= \alpha_2 z_0^4 - D_x x \\
\frac{dy}{dt} &= k_2 x - (D_y + \alpha_3) y,
\end{aligned} \tag{5.4}$$

5.4.2 Results and discussion

Table 5.1 shows the parameter value estimates used in this analysis. The half-life of p53 is between 20 and 60 mins in unstressed cells with the majority of estimates placing it at around 30 mins (Reich *et al.*, 1983; Maki and Howley, 1997; Vierboom *et al.*, 2000;

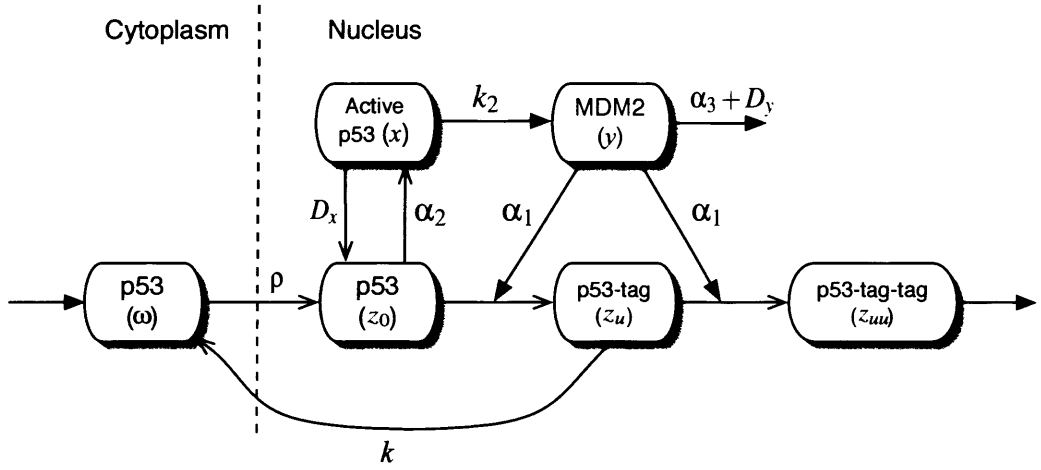


Figure 5.3: A schematic of a p53 localisation model with active p53 and MDM2 (the square pulse model).

Friedler *et al.*, 2003). Filippini *et al.* measured, through flow cytometry, that the NALM-6 cell line, which contains wild-type p53, contains approximately 10,000 p53 molecules ($\approx 1.66 \times 10^{-20}$ mol) (Filippini *et al.*, 1998). Using this information and a simple model of the turnover of p53, p_z was estimated. MDM2 has been found to have a half-life between 15 and 30 mins and so a half-life of 20 mins was used to estimate D_y (Teoh *et al.*, 1997). The rest of the parameters were estimated based on the above information and the following assumptions,

- There is approximately ten times more cytoplasmic p53 than there is nuclear p53. It is known that when the cell is not stressed, the majority of the p53 is in the cytoplasm, this means that p53 export must be stronger than input and so the export rate (k) needs to be high.
- There are approximately four times less active p53 tetramers than z_0 i.e. there is an equal number of p53 molecules in both states. This is because when a cell is not stressed there should be a minimal amount of active p53 or apoptotic effects might be produced.
- Once p53 is in the double ubiquitinated state it is quickly degraded. This means that β is set high.
- It was arbitrarily chosen to have the amount of MDM2 at equilibrium set at approximately 2×10^{-20} mol.

It is assumed that there is no active ATM in unstressed cells and so $\alpha_3 = 0$. The majority of the parameter values seem to take reasonable values (Cinquin, 2006) but the p53 activation rate seems high when considering that the rate of protein-protein association is considered to be between 10^6 and 10^9 mol^{-1} s^{-1} (Northrup and Erickson,

Table 5.1: The parameter values used in the analysis of the square pulse model.

Parameter	Value
ρ	$1.5 \times 10^{-3}\ s^{-1}$
p_z	6.5×10^{-24} mol s^{-1}
α_1	5.96×10^{17} mol^{-1} s^{-1}
α_2	5×10^{58} mol^{-3} s^{-1}
D_x	$1 \times 10^{-3}\ s^{-1}$
k	$2.6 \times 10^{-2}\ s^{-1}$
β	$5.2 \times 10^{-2}\ s^{-1}$
α_3	$0\ s^{-1}$
D_y	$5.8 \times 10^{-4}\ s^{-1}$
k_2	0.027 mol^{-1} s^{-1}

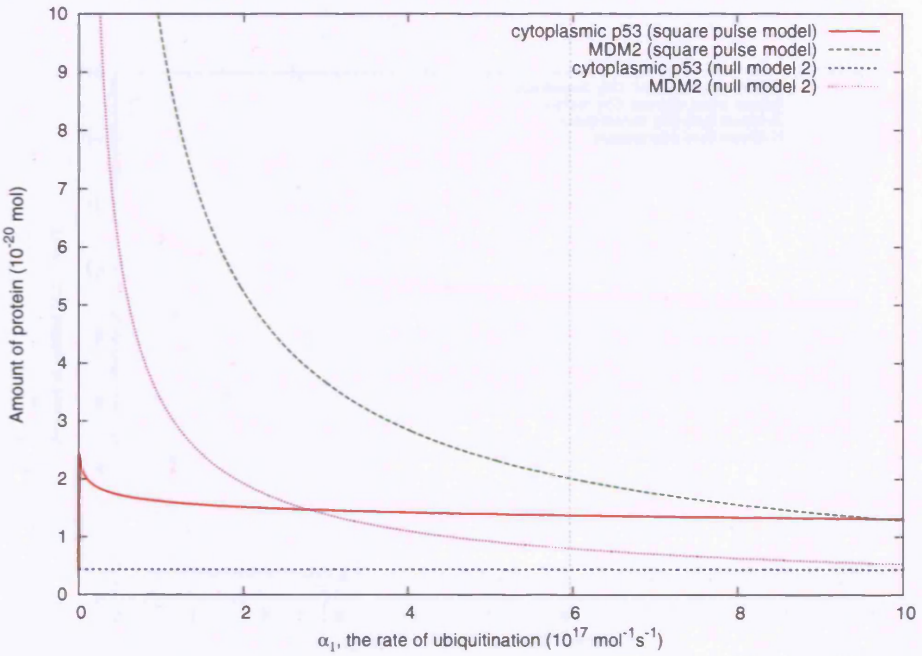
Table 5.2: The equilibrium condition for the square pulse model and null model 2.

Component	Square pulse model	Null model 2
ω	1.372×10^{-20}	4.333×10^{-21}
z_0	1.715×10^{-21}	1.362×10^{-21}
z_u	5.417×10^{-22}	–
x	4.328×10^{-22}	1.713×10^{-22}
y	2.015×10^{-20}	8.002×10^{-21}

1992; Gabdoulline and Wade, 1997). These rates of protein-protein association are based on diffusing molecules in a liquid with no structure whereas in the cell the construction of the tetramer occurs around DNA which will be a more active process, so a faster rate could be possible.

As was designed, the amount of cytoplasmic p53 at equilibrium when the example parameters are used is greater than the amount of p53 in the nucleus and the total amount of p53 (1.653×10^{-20} mol) is approximately correct (Table 5.2). There is generally significantly less of each component at equilibrium for the null model because p53 is not re-circulated. The equilibrium of most of the components follow a similar pattern as α_1 (the rate of ubiquitination by MDM2) is varied with the other parameters set at the example values (Figure 5.4). The exception is cytoplasmic p53 which stays at a constant level. The equilibrium levels are always higher in the square pulse model than in null model 2, because the square pulse model re-circulates some ubiquitinated p53 whereas in null model 2 it is all destroyed. As α_1 is decreased the difference between the square pulse model and the null model increases, indicating that it is only when the cell is stressed that the differences between the two models will be significant. The greatest effect of altering the value of α_1 is on MDM2 and active p53 suggesting that inhibiting MDM2 ubiquitination is an effective mechanism to increase levels of active p53.

As α_2 (the rate of p53 conversion to its tetramer form) is increased (Figure 5.5), the equilibrium level of z_0 decreases for both models and the level of ω decreases for the square pulse model. MDM2 and active p53 have a corresponding increase. This is



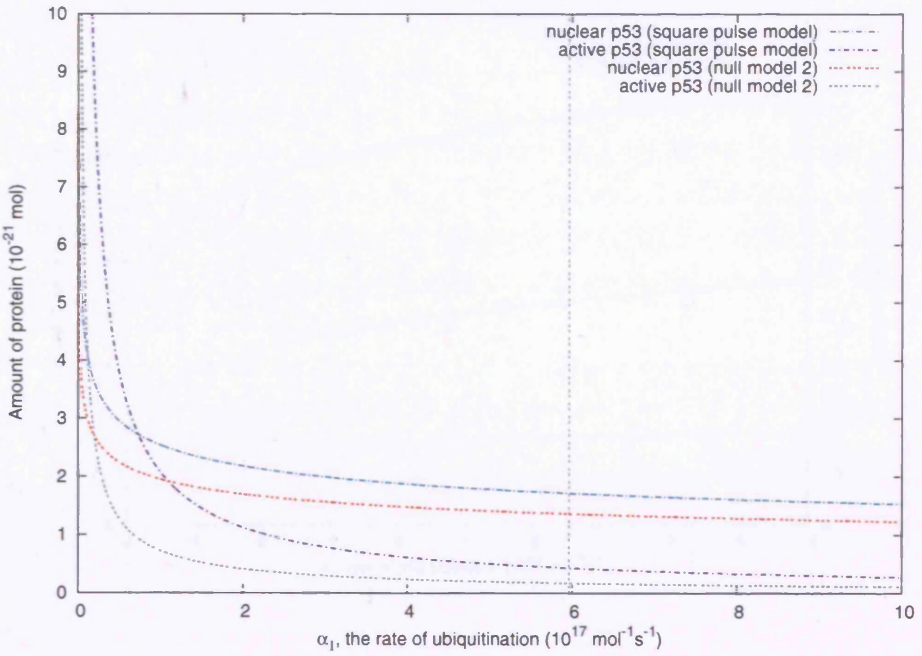


Figure 5.4: A plot to show the variation of the equilibrium condition with the rate of ubiquitination (α_1) for the square pulse model (equation 5.3) and null model 2 (equation 5.4). The vertical line indicates the example parameter set value, $\alpha_1 = 5.96 \times 10^{17}$ mol^{-1} s^{-1}.

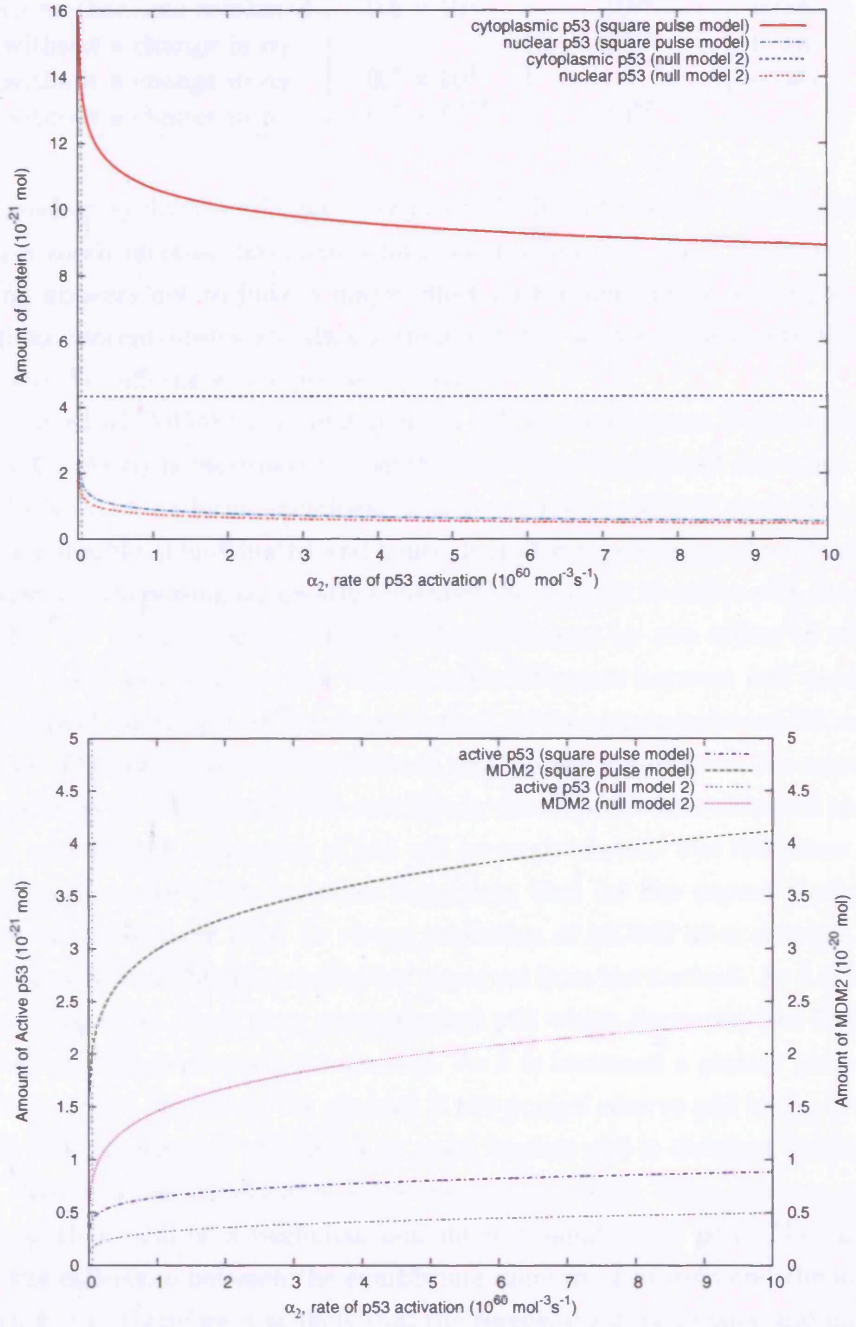


Figure 5.5: A plot to show the variation of the equilibrium condition with the rate of p53 activation (α_2) for the square pulse model (equation 5.3) and null model 2 (equation 5.4). The vertical line indicates the example parameter set value, $\alpha_2 = 5 \times 10^{58}$ mol^{-3} s^{-1}.

Table 5.3: The parameter values that are changed to replicate a square pulse of DNA damage

	α_1 mol^{-1} s^{-1}	α_2 mol^{-3} s^{-1}	α_3 s^{-1}
All mechanisms combined	0.5×10^{17}	10^{64}	0.005
without a change in α_1	-	10^{64}	0.005
without a change in α_2	0.5×10^{17}	-	0.005
without a change in α_3	0.5×10^{17}	10^{64}	-

because increasing α_2 diverts a larger amount of p53 from the chain mechanism. When α_2 is small, a small increase produces a large change but the effect decreases as α_2 is increased. α_2 appears not to have a major effect on the amount of active p53. Again, the equilibrium concentrations are always greater in the square pulse model than in the null model and the difference changes as α_2 varies.

The rate at which MDM2 is inhibited when cell damage occurs is controlled by α_3 (see Figure 5.6). As α_3 is increased the equilibrium value for MDM2 decreases (in both models) and p53 in all its forms increases. This is because as MDM2 decreases, the rate at which p53 is double ubiquitinated and hence degraded is reduced and so there is more p53 in the system. Increasing α_3 greatly enhances the amount of active p53, for example when $\alpha_3 = 0.1\ s^{-1}$ the amount of active p53 has increased by two orders of magnitude from its $\alpha_3 = 0\ s^{-1}$ value. There is a considerable difference between null model 2 and square pulse model, with α_3 having a greater effect on the square pulse model, especially for active p53. The reason that the amount of p53 increases quicker in the square pulse model is because as α_3 is increased, it is more likely that if p53 is ubiquitinated that it will be exported, so a greater proportion of p53 will be re-circulated. The difference between the two models increases as α_3 increases suggesting that for the export mechanism to play an important role there must be strong inhibition of MDM2 after damage.

k is the rate at which ubiquitinated p53 is exported from the nucleus. As k is increased all components increase apart from ubiquitinated p53 which decreases (see Figure 5.7). The rate of change decreases as k is increased. As k is increased a greater proportion of tagged p53 is recycled increasing the amount in the pool of reserve p53 in the cytoplasm. The amount of ubiquitinated p53 drops as more nuclear p53 is diverted leaving less in that state. When a cell experiences DNA damage k will in effect drop to near zero values because there will be a negligible amount of ubiquitinated p53. The larger k is, the greater the difference between the equilibrium amount of protein and the amount of protein when $k = 0$, therefore it is likely that the response will be greater and more rapid when k is higher.

A square pulse of DNA damage can be simulated in this model by changing the α parameter values in the model and then changing them back to their initial values after a certain time. This will give some idea of the dynamics of the square pulse model after DNA damage is experienced. Formally, the αs are not parameters any more but variables.

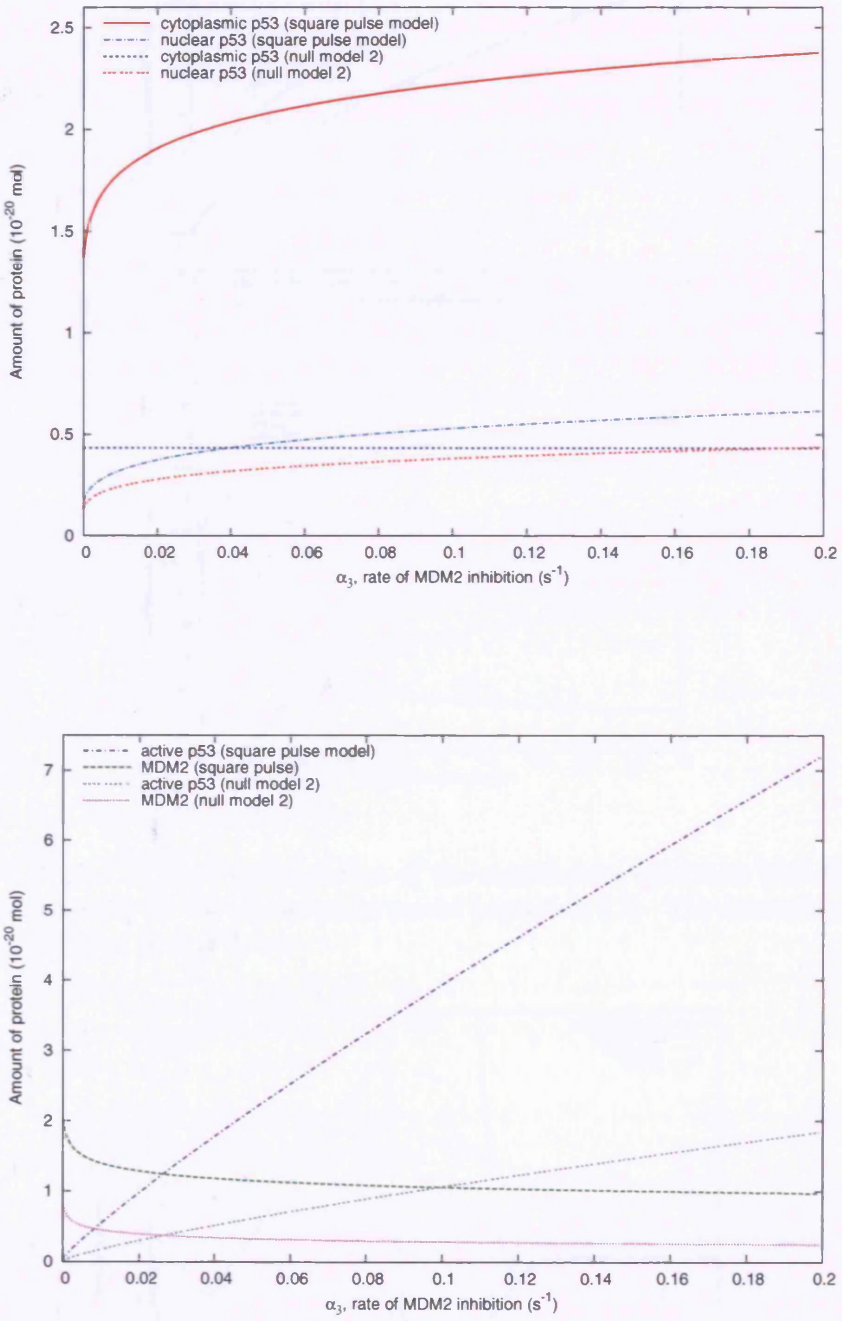


Figure 5.6: A plot to show the variation of the equilibrium condition with the rate of MDM2 inhibition (α_3) for the square pulse model (equation 5.3) and null model 2 (equation 5.4). The example parameter set value is $\alpha_3 = 0$.

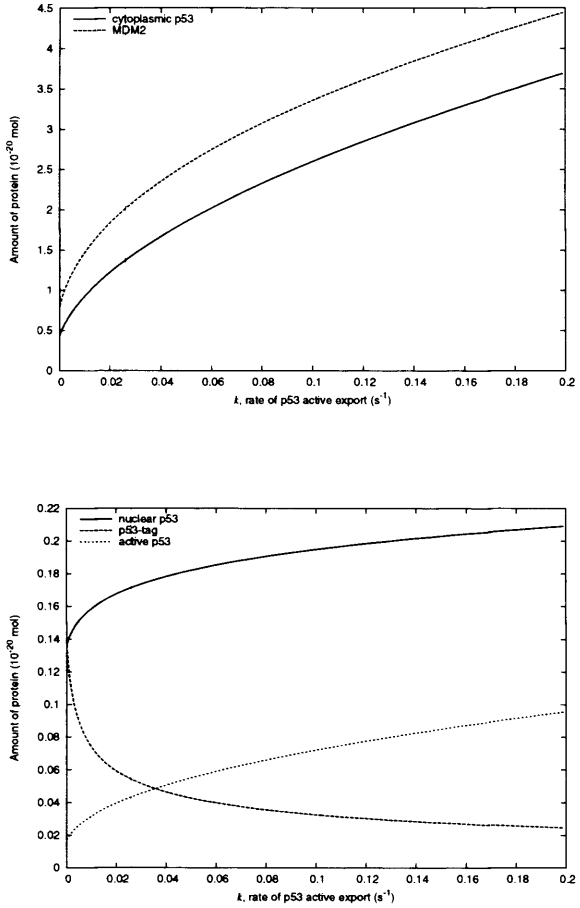


Figure 5.7: A plot to show the variation of the equilibrium condition with the rate of p53 nuclear export (k) for the square pulse model (equation 5.3). The example parameter set value is $k = 2.6 \times 10^{-2}\ s^{-1}$.

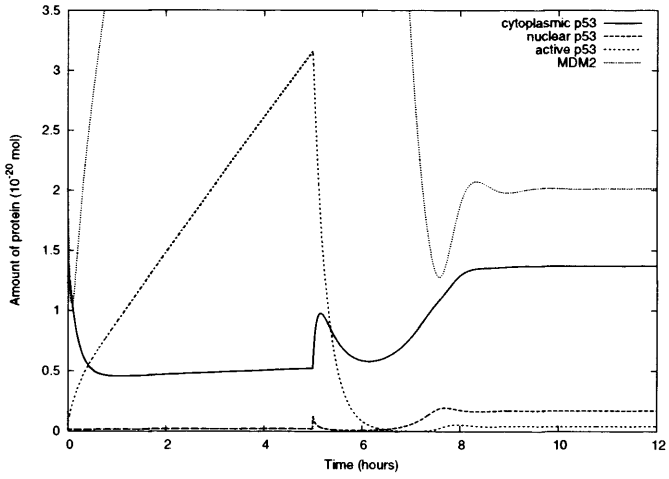


Figure 5.8: A plot to show the reaction of the square pulse model (equation 5.3) to 5 hours of DNA damage all mechanisms controlled by α_1, α_2 & α_3 are changed.

A number of different parameter changes will be made to test the different mechanisms of the model. Table 5.3 summarises these changes, after 5 hours the parameter values will revert to their original values.

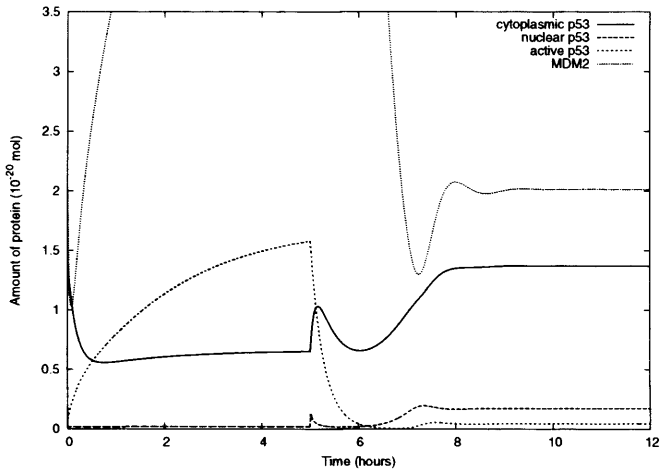
The square pulse model has a fairly similar set of dynamics in each of the situations tested (Figures 5.8 & 5.9). It takes a considerable amount of time to return to equilibrium after the damage has been repaired ($\approx$ 5 hours), but it is still biologically reasonable. The dynamics show damped oscillations occurring between p53 and MDM2 during the recovery which would agree with experimental observations (Bar-Or *et al.*, 2000). This is encouraging and hints that the models are capturing at least some of the dynamics appropriately.

The level of active p53 is the most important feature of the response as it triggers the apoptotic and DNA repair mechanisms. In all of the situations the amount of active p53 rises rapidly and the majority of the p53 moves into the nucleus as expected (Figure 5.10(a)). The quickest rise in active p53 occurs when all of the mechanisms are triggered, when any of the mechanisms are not used the performance is significantly reduced. All of the mechanisms seem to have an approximately equal effect on the response. When α_2 remains constant there is only slightly more p53 in the nucleus than in the cytoplasm (see Figure 5.9(b)), suggesting that an increased rate of activation is required to get the observed accumulation of p53 in the nucleus.

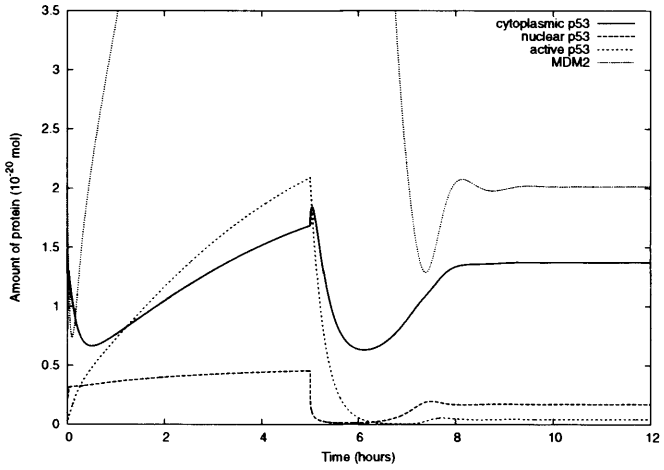
In all situations the level of MDM2 rises after DNA damage (Figure 5.10(b)). A similar dynamic is followed apart from when there is no increase in the degradation rate of MDM2 (α_3 remains constant); MDM2 rises but at a suppressed rate given the level of active p53, when the damage signal is switched off there is a very large pulse of MDM2 (this might occur because the production rate of MDM2 does not saturate). When α_3 remains constant MDM2 rises at a more rapid rate as there is no mechanism to restrict the level of MDM2. It is unclear whether MDM2 should rise after DNA damage as protein data suggests it will rise or fall depending on the level of damage (see section 3.4.3). The dynamics of cytoplasmic p53 differ depending on which mechanisms are activated by DNA damage but it is generally initially suppressed. In all cases the levels of nuclear p53 are suppressed during the damage signal apart from when α_2 remains constant, and in this case the amount of nuclear p53 slowly rises.

The overall dynamics of null model 2 when all the mechanisms are activated are the same as in the square pulse model (Figure 5.11), but the amplitudes are different and the level of cytoplasmic p53 remains constant. In all three situations the response of active p53 is faster and the peaks higher in the square pulse model (Figure 5.12). This suggests that the export mechanism causes an improvement in the response of the p53 network to damage. The time it takes to get back to equilibrium is similar for both models. The differences are due to the "pool" of cytoplasmic p53 that exists when there is an export mechanism; when there is damage this "pool" rushes into the nucleus. Also this has the effect that more p53 is re-circulated for the square pulse model, even when there is a

(a)



(b)



(c)

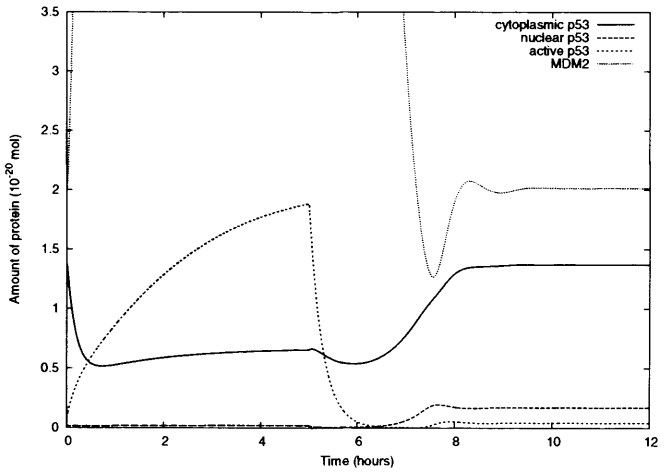
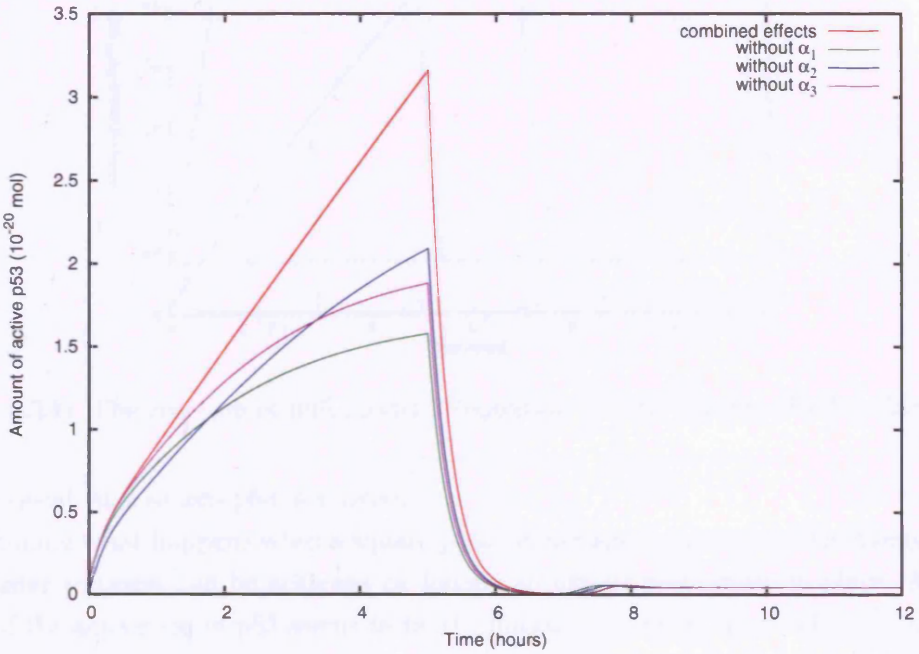


Figure 5.9: A plot to show the reaction of the square pulse model (equation 5.3) to 5 hours of DNA damage with (a) α_1 remaining unchanged, (b) α_2 remaining unchanged and (c) α_3 remaining unchanged. See Table 5.3 for mechanisms.

(a)



(b)

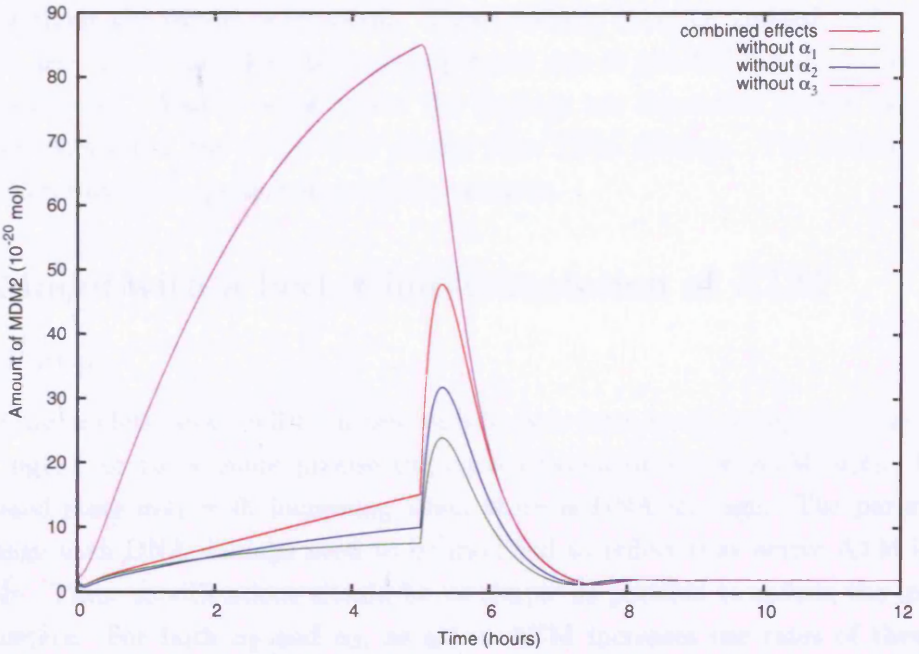


Figure 5.10: A plot to show how the different mechanisms of the square pulse model (equation 5.3) affect (a) active p53 and (b) MDM2.

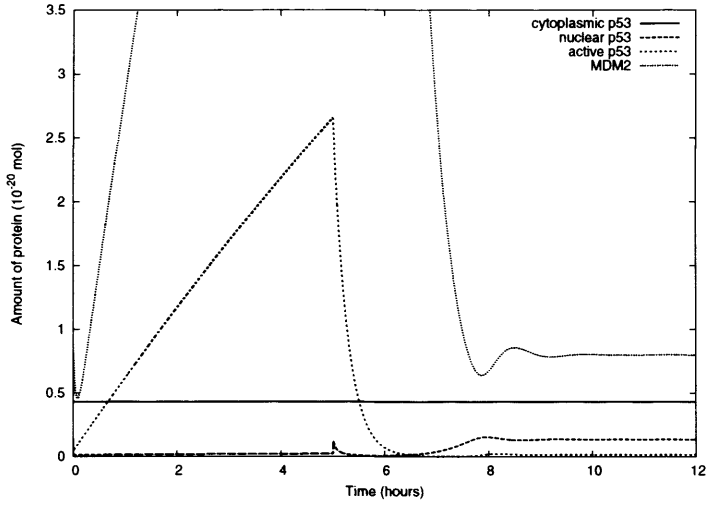


Figure 5.11: The reaction of null model 2 (equation 5.4) to 5 hours of DNA damage.

damage signal, and so less p53 is wasted.

Examining what happens when a square pulse of damage occurs in a cell has suggested that a faster response can be achieved by having an export mechanism in place. ATM's control of the activation of p53 seems to be the major mechanism by which the majority of p53 is retained in the nucleus. ATM's control of MDM2 inactivation and ubiquitination allows p53 to be recycled without being destroyed and this feeds the response. Despite these improvements it appears that the cell can get back to equilibrium in just as rapid time as without the export mechanism. These findings must be treated with caution though as the effects are likely to be exaggerated due to the DNA damage occurring as a square pulse. Also to some extent the findings are dependent on the estimated parameter values and how much they change after DNA damage. The extent of this could be determined by performing stability analysis.

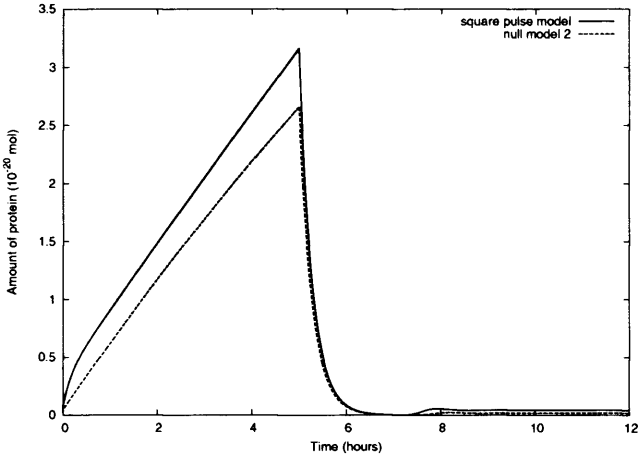
5.5 Model with a better implementation of ATM

5.5.1 Setup

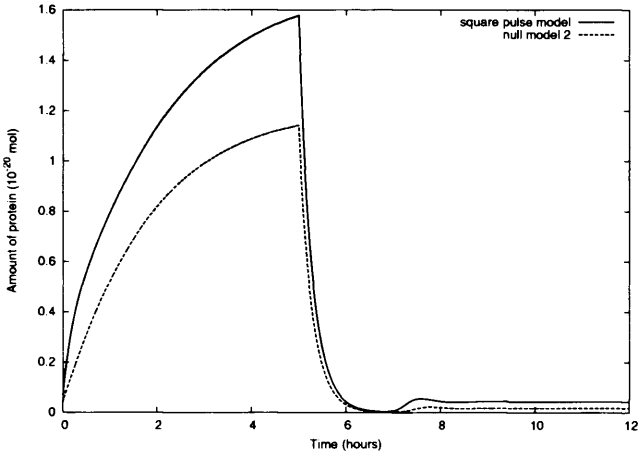
To make the models more realistic a new variable was introduced to represent the DNA damage signal, or to be more precise the concentration of active ATM, $a(t)$. In the non-stressed state $a(t) = 0$, increasing when there is DNA damage. The parameters that change with DNA damage need to be modified to reflect that active ATM is now a variable. These modifications should be as simple as possible to reduce the number of parameters. For both α_2 and α_3, as active ATM increases the rates of these two mechanisms increase so the following modifications are made,

$$\begin{aligned} \alpha_2^{old} &= \alpha_2^{basal} + \alpha_2 a, \\ \alpha_3^{old} &= \alpha_3 a. \end{aligned}$$

(a)



(b)



(c)

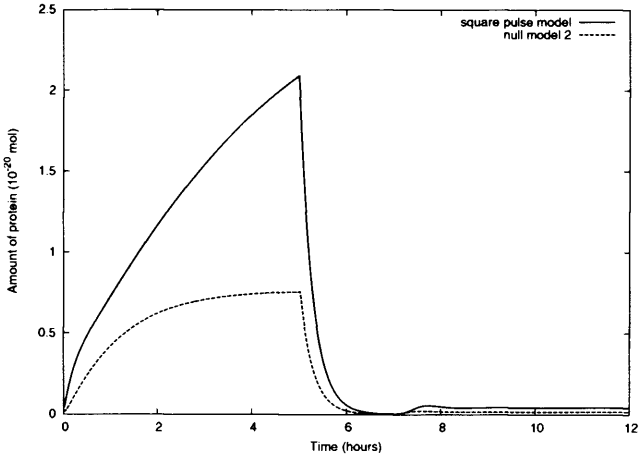


Figure 5.12: A comparison of the dynamics of active p53 in the square pulse model (equation 5.3) and null model 2 (equation 5.4) after 5 hours of DNA damage in the following situations (a) combined effect (b) without change in α_2 mechanism (c) without change in α_3 mechanism.

α_2^{basal} is introduced as there will be some slow rate of activation of p53 even when there is no damage. As active ATM is increased, the rate of ubiquitination of p53 is suppressed. A simple mathematical form that represents this is,

$$\alpha_1^{old} = \frac{\mu}{a+\kappa}.$$

Therefore, the model becomes,

$$\begin{aligned}
\frac{d\omega}{dt} &= p_z + kz_u - \rho\omega, \\
\frac{dz_0}{dt} &= \rho\omega - \frac{\mu}{a+\kappa} z_0 y - 4(\alpha_2^{basal} + \alpha_2 a) z_0^4 + 4D_x x, \\
\frac{dz_u}{dt} &= \frac{\mu}{a+\kappa} z_0 y - kz_u - \frac{\mu}{a+\kappa} z_u y, \\
\frac{dz_{uu}}{dt} &= \frac{\mu}{a+\kappa} z_u y - \beta z_{uu}, \\
\frac{dx}{dt} &= (\alpha_2^{basal} + \alpha_2 a) z_0^4 - D_x x, \\
\frac{dy}{dt} &= k_2 x - (D_y + \alpha_3 a) y, \\
a &= a(t).
\end{aligned} \tag{5.5}$$

This model is called the *full model.* The corresponding null model (*null model 3*) is,

$$\begin{aligned}
\frac{d\omega}{dt} &= p_z - \rho\omega, \\
\frac{dz_0}{dt} &= \rho\omega - \frac{\mu}{a+\kappa} z_0 y - 4(\alpha_2^{basal} + \alpha_2 a) z_0^4 + 4D_x x, \\
\frac{dx}{dt} &= (\alpha_2^{basal} + \alpha_2 a) z_0^4 - D_x x, \\
\frac{dy}{dt} &= k_2 x - (D_y + \alpha_3 a) y, \\
a &= a(t).
\end{aligned} \tag{5.6}$$

In an individual cell, after DNA damage there will be multiple strand breaks, each "emitting" a certain amount of signal. When there is 0.5Gy of damage there are approximately 20 strand breaks (see section 3.3) which will mean the signal will drop in steps as the breaks are repaired. Here though it is assumed that by the time the signal is represented by active ATM it will be smoothed enough to be represented by an exponential i.e.

$$a(t) = a_0 e^{-1.39\times10^{-4}t},$$

the degradation rate constant of $1.39 \times 10^{-4}\ s^{-1}$ comes from experimental data (see section 3.3). The radiation dose is assumed to be proportional to the amount of active ATM. Active ATM is considered to be in Gy units.

5.5.2 Results and discussion

Apart from the α constants the same example parameters were used as in the previous section. α_1 should be reasonably high when there is little or no damage and be zero when there is a lot of damage i.e. when $a = 5$, $\alpha_1 = 0.5 \times 10^{17}$ mol^{-1} s^{-1} and when $a = 0.01$ then $\alpha_1 = 5.96 \times 10^{17}$ mol^{-1} s^{-1}. This gives the parameters $\kappa = 0.44696$ and $\mu = 2.72348 \times 10^{17}$. $\alpha_2^{basal} = 5 \times 10^{58}$ mol^{-3} s^{-1} a low rate equal to the value of α_2 in the previous section. It is necessary for active ATM to have a large effect on the activation of p53 and the inhibition of p53 and so $\alpha_2 = 2 \times 10^{64}$ mol^{-3} s^{-1} and $\alpha_3 = 0.01$ s^{-1}.

Table 5.4: The equilibrium conditions for the full model and the null model.

Component	Square pulse model	Null model
ω	1.370×10^{-20}	4.333×10^{-21}
z_0	1.708×10^{-21}	1.357×10^{-21}
z_u	5.404×10^{-22}	–
x	4.258×10^{-22}	1.696×10^{-22}
y	1.982×10^{-20}	7.893×10^{-21}

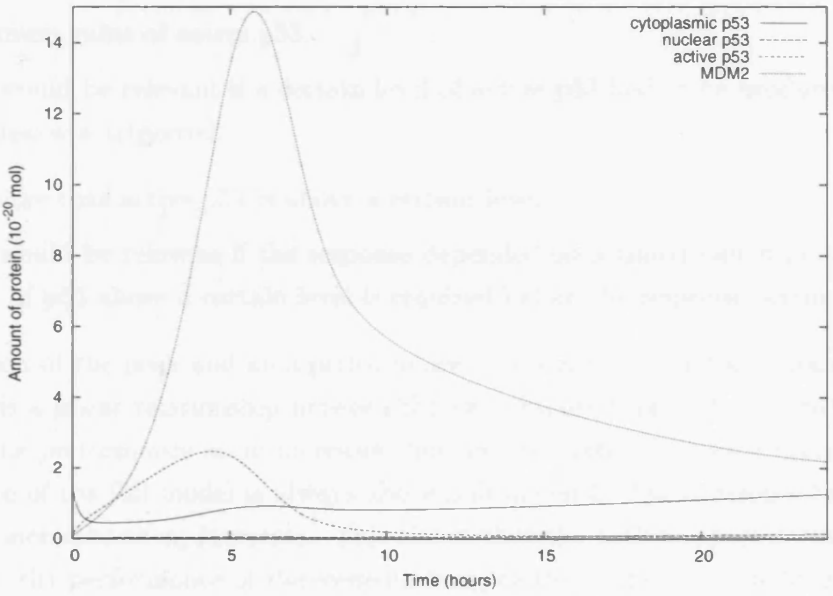
As designed, the equilibrium conditions when $a_0 = 0$ are very similar to those for the square pulse model (Table 5.4 & Table 5.2). They will be used as initial conditions for the rest of the experiments. Example runs when $a_0 = 5$Gy (a high level of damage) show dynamics that are similar to the square pulse model dynamics but the behaviour is smoother and the peaks of active p53 are less extreme (Figure 5.13). The key features remain; there is a rapid accumulation of active p53 resulting in the majority of p53 being in the nucleus. Unlike the square pulse model damped oscillations are not seen except possibly in cytoplasmic and nuclear p53. There is a considerably longer recovery time than the square pulse model (approximately 30 hours), which is probably because MDM2 does not reach such large values. The active transport mechanism still provides an improved reaction to damage. The peak in active p53 occurs at about 5 hours, which is within the biologically expected range (see section 3.4.2).

Examining example runs provides some insight into the system but it is also worth examining how the various mechanisms under study affect the performance of the system. The main output of the system that will propagate the damage signal will be the level of active p53. It is not known how the level of p53 affects the processes downstream of it and so three performance scores will be defined that measure the strength of the response:

1. The amount of extra active p53 gained from the signal.

   This score would be relevant if the "switch" to start apoptosis depends on the total amount of extra active p53 that is produced over the response, rather than its value at a particular moment i.e. the response can be short and sharp or long and shallow. This is calculated by approximating the area under the active p53

(a)



(b)

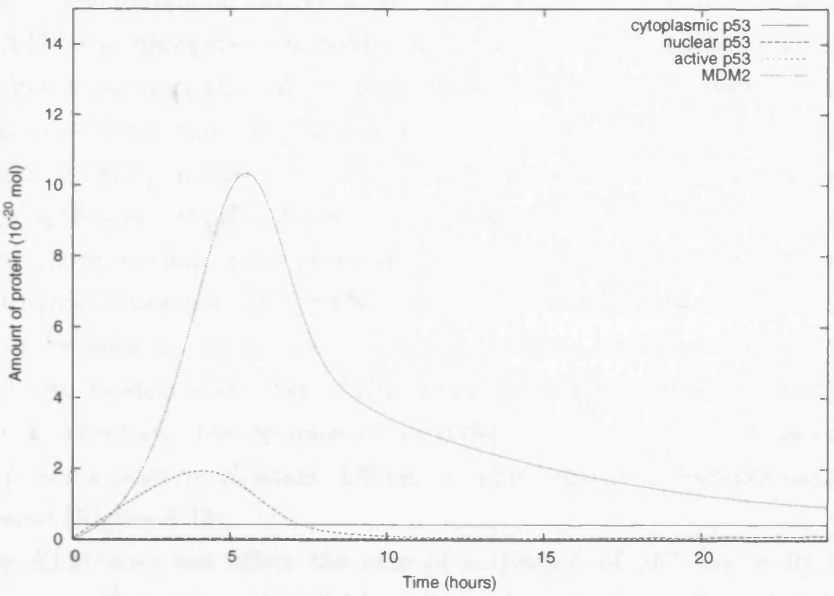


Figure 5.13: Examples of the response of the system to DNA damage when $a_0 = 5$ for (a) the full model (equation 5.5)and (b) the corresponding null model (equation 5.6).

minus equilibrium value curve. Negative areas are ignored. This is implemented by joining two adjacent points by a straight line and calculating the area under that line, if the next point is negative this is not added to the sum (Simpson's rule was not used as the data points are not equally spaced).

2. Maximum value of active p53.

   This would be relevant if a certain level of active p53 had to be produced before a response was triggered.

3. The time that active p53 is above a certain level.

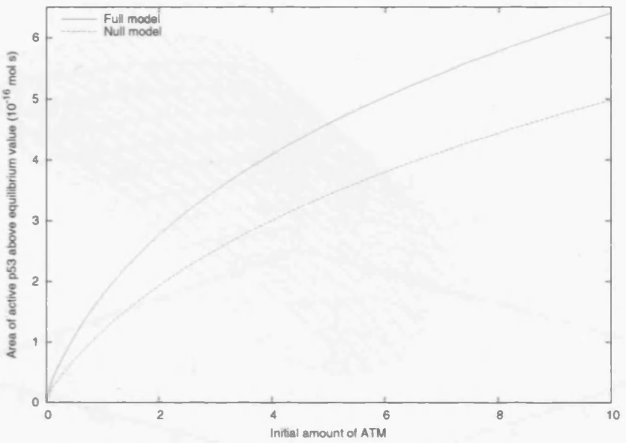
   This would be relevant if the response depended on a timed switch i.e. a constant input of p53 above a certain level is required before the response occurs.

The variation of the peak and area performance scores are very similar, making it likely that there is a linear relationship between the two (Figure 5.14(a) & 5.14(b)). As a_0 is increased the performance score increases, but also the rate of increase decreases. The performance of the full model is always above null model 3. The difference between the two scores increases as a_0 increases. This shows that the active export mechanism has an effect on the performance of the system. Imagine there was a threshold that needed to be passed before the apoptosis machinery was activated, the higher the threshold is the greater the difference in the amount of DNA damage that is needed to trigger it. For example, if the threshold was set at an area of 4.5×10^{-16} mol s^{-1} then the initial amount of ATM and hence damage needed to activate the full model ($\approx$ 4.8 Gy) is 60% of that required to activate the null model ($\approx$ 8 Gy). This is an impressive improvement.

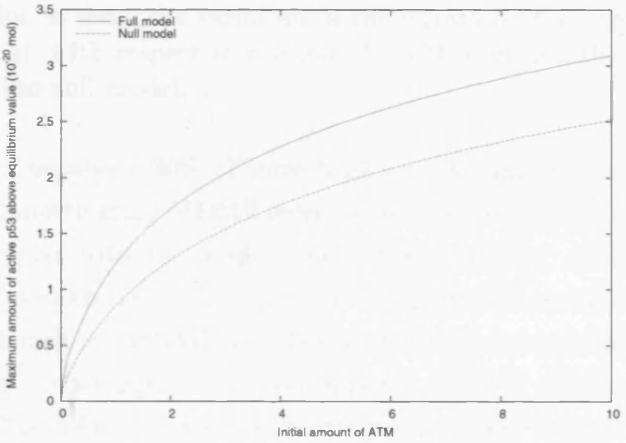
The variation of the time duration performance score has the same shape for both models (Figure 5.14(c)). Initially, the duration remains at zero until the damage signal is large enough so that active p53 can overcome the threshold. After this, the duration rises very rapidly (more so than the other scores) but then the improvement starts to slow down. Null model 3 reaches the threshold at a larger amount of damage than the full model and the performance score always remains less than the full model. The difference between the two models starts very high (when null model score begins to rise) but decreases as a_0 increases. The decrease in the difference slows as a_0 increases and it may be tending to some positive constant difference. This behaviour does not depend on the threshold value (Figure 5.15).

If active ATM does not affect the rate of activation of p53 ($\alpha_2 = 0$) the performance curves are the same as if it did but the scale is dramatically reduced, dropping by approximately a third (Figure 5.16(b)). This suggests that ATM increasing the rate of activation of p53 is essential to produce a strong response. An interesting effect of knocking out ATM activation of p53 is that the difference between the null and full model is greater. This occurs because once p53 is activated it is completely removed from the ubiquitination mechanism. When ATM's inhibition of ubiquitination is knocked out the

(a)



(b)



(c)

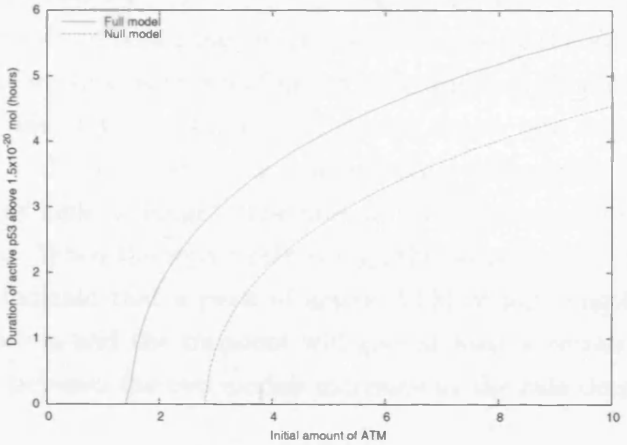


Figure 5.14: A plot showing the variation of the strength of response with the initial amount of damage. Three performance scores are shown (a) total amount of p53 above equilibrium value, (b) the maximum amount of active p53 and (c) the duration of active p53 at levels above 1.5×10^{-20} mol.

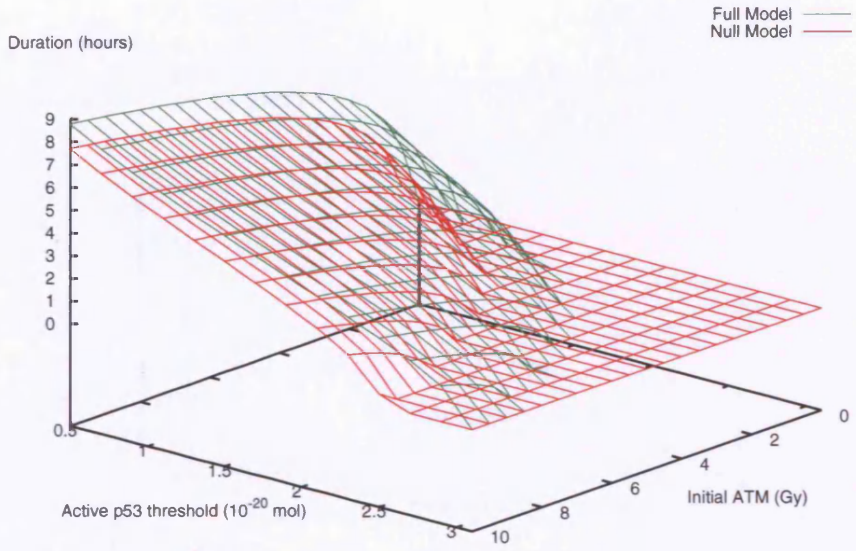


Figure 5.15: A plot to show the variation of the duration of active p53 concentration above some threshold, with respect to the initial ATM level and the threshold for both the full model and the null model.

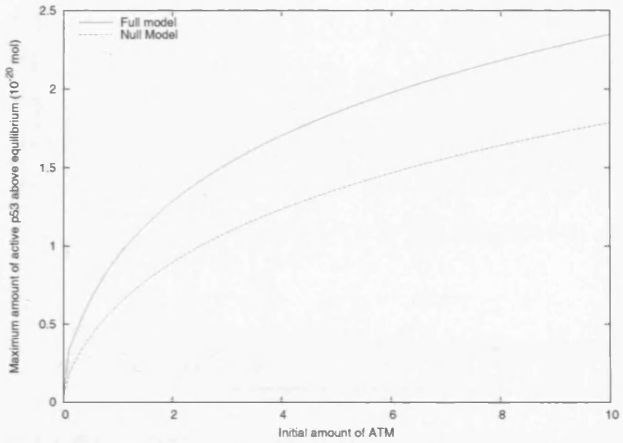
score is scaled down by about 20% (Figure 5.16(a)). A similar effect is seen when the mechanism of ATM inactivating MDM2 levels is removed but here the reduction in performance is much larger with the score being reduced by over a half (Figure 5.16(c)). This indicates that even though it is helpful to disrupt the ubiquitination mechanism, it is more effective to knock out MDM2 completely. All three mechanisms produce substantial improvements to the strength of the p53 response to DNA damage, but the complete inhibition of MDM2 is the most effective mechanism to improve performance as without MDM2 there is no inhibition of p53 at all.

The rate of repair used was based on experimental data, but it would be interesting to know what effect rate of repair has on the performance of the system (Figure 5.17). A repair rate of 0.1 corresponds to a half-life of approximately 7 hours and a rate of 2 is equivalent to a half life of about 20 minutes. As the repair rate decreases the strength of response increases, this is because the damage signal will persist and so the amount of active p53 will stay high for longer. The rate of change increases rapidly as the repair rate approaches zero. When the repair rate is high the score is at a low almost constant value. This might indicate that a peak of active ATM of any length will displace the system from equilibrium and the transient will give at least a certain amount of active p53. The difference between the two models increases as the rate decreases.

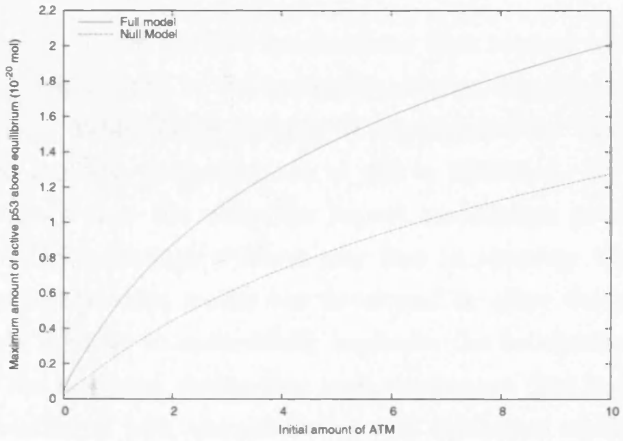
5.6 Conclusion

In this chapter a number of different models have been proposed and examined to study the effect that localisation mechanisms have on the way that the p53 gene network system

(a)



(b)



(c)

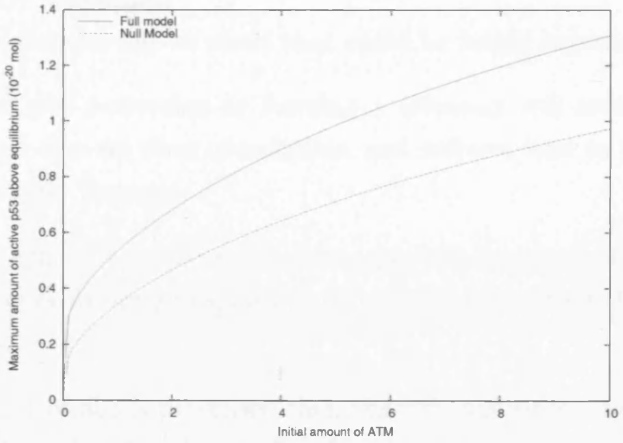


Figure 5.16: A plot to show the effect on the sensitivity curve of knocking out different mechanisms: (a) the inhibition of ubiquitination by ATM (b) the activation of p53 by ATM and (c) the inhibition of MDM2 by ATM.

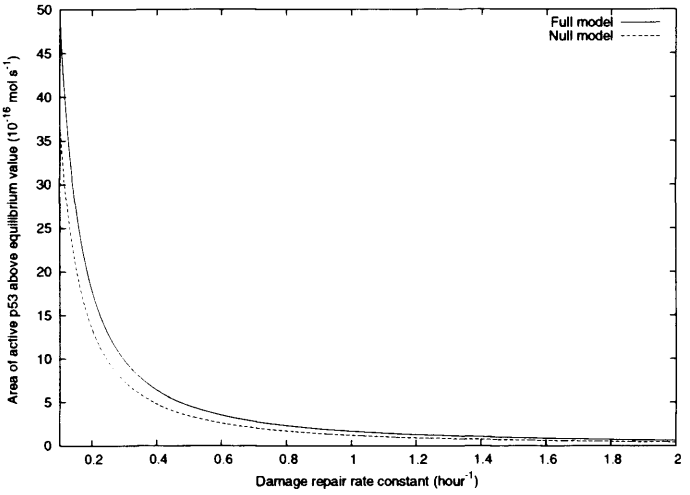


Figure 5.17: A plot to show the variation of the total amount of extra active p53 with the repair rate, for both the full model and the null model. a_0 was set to 5Gy.

reacts to DNA damage. There are two mechanisms that control where p53 is located: the activation of p53 (retains p53 in the nucleus) and the ubiquitination of p53 (exports p53 from the nucleus). When DNA damage is experienced by the cell the activation of p53 is encouraged and the ubiquitination of p53 is inhibited. By comparing with a null model it was shown that the ubiquitin export mechanism produced a faster and stronger reaction to DNA damage without any loss in recovery time. This suggests that the two state ubiquitination model has developed to allow the efficient reaction to damage. The models manage to successfully replicate the localisation of p53 with and without damage. It has also been shown that each mechanism that is altered by the DNA damage signal (activation of p53, ubiquitination and inhibition of MDM2) significantly contributes to the performance of the reaction but it seems likely that the inhibition of MDM2 is particularly important.

A number of predictions can be made that could be tested experimentally:

1. Preventing the p53 activating or forming a tetramer will reduce the amount of apoptosis found at a set dose of radiation and will not lead to p53 being retained in the nucleus after damage.

2. In cancer cells where localisation is used to prevent apoptosis, the ATM-p53 interaction or tetramer-forming mechanism is mutated preventing the retention of p53 in the nucleus.

3. If nuclear export of p53 is prevented then the cells will be less sensitive to damage, there will be less phosphorylated p53 after damage, and phosphorylated p53 will increase at a slower rate and at equilibrium there will be a more even distribution of p53 between nucleus and cytoplasm.

The localisation model that was used here had many simplifications. Even though the

model replicated the known behaviour well in a broad way there were discrepancies; the recovery time was too long and MDM2 does not follow the correct dynamics (see section 3.4.3). There were no oscillations between MDM2 and p53 but there is debate about whether these occur at high levels of damage. Improvements might be made by modifying the model to include more complexity, for example one could:

- Include a more accurate model of p53 activation. This would have to include the fact that dimers are formed before tetramers and that tetramers are formed on the DNA.
- Allow the ubiquitination rate from p53 to p53-tag and p53-tag to p53-tag-tag to be different. For example, if only one tag is needed for export and 6 tags is needed for degradation then really the rate from p53-tag to p53-tag-tag should be 1/5 of the rate from p53 to p53-tag.
- Allow p53 to be ubiquitinated in the nucleus and then de-ubiquitinated at a certain rate.
- Include MDM2 localisation properties.
- Include additional components of model such as ARF and E2F1.
- Develop a more detailed model of ubiquitination, including multiple sites.

It would be interesting to model the two proposed methods for p53 export i.e. MDM2 ubiquitinating p53 or MDM2 binding p53 and both shuttling out of the nucleus. One could ask the question of which method produces the most "efficient" results. It would also be worth examining whether different levels of damage produce different dynamics. Finally, it would be useful to examine whether the model can replicate some of the other experiments performed by Li *et al.* (2003) to test the model further.

Ciliberto *et al.* (2005) have produced a model of the p53 network that replicates the pulse like dynamics of p53 (see section 2.4.4). The localisation of the components is a key part of this model but unlike the models developed here it is the location of MDM2 that is considered rather than p53. Despite this, it emphasises that nuclear localisation plays an important role in this system. Unfortunately this work was published too late to affect this analysis.

Throughout this chapter the results have been basically qualitative because the parameter values can only be estimated approximately. To go further the parameter values either need to be measured directly (which can be time consuming and costly) or found through parameter estimation. This would allow more biological conclusions to be made and the formal comparison of models. Parameter estimation will be examined in the next few chapters.

Chapter 6

Parameter estimation for a mathematical model of the p53 protein network using established methods

6.1 Introduction

After models have been proposed it is important to know whether the model behaves in same way as the system that it represents. Without some confidence in the model it is impossible to make conclusions and predictions. One necessary criteria for a model to be useful is that the results it produces are in agreement with experimental data (the model should also reveal something new about the system under study and not be so complicated it over-fits the data). Therefore, a quantitative measure is needed to measure the distance between the model results and the experimental data; this measure is called an error function (Press *et al.*, 2002).

If a model's parameter values are known, then the model will have a set behaviour and can be easily compared with the data. Normally though, some if not all parameter values are unknown. This is because the values vary depending on the situation (for example, between cell types) and experiments to determine them are complex. Depending on the parameter values, the model will have different dynamics. Finding the parameter values that produce the smallest error function value is known as parameter estimation and is the most difficult part of the modelling process (Tyson, 1999). If the best possible error function value is high then the model is probably not a good representation of the system, but if the error function value is small then it is *possible* that the model can represent the system. Also if the parameter values are unrealistic then the model can be discarded. Parameter estimates can also provide useful information about the relative importance of mechanisms in the system.

An error function is also necessary for the comparison of models. If a model has a worse error function value than another (taking into account the number of parameters in each model), then it can be rejected. There is also the potential for network building through the comparison of error functions. For example, if a relationship between two proteins is added to the model and this fits the data considerably better than without the relationship it suggests that this relationship may exist in the biological system. In practice, it is normally necessary to perform parameter estimation before the comparison can take place, and it is common to use the same error function for both the estimation and the comparison.

6.1.1 The parameter estimation problem

A general ODE model can be defined as follows:

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{f}(\mathbf{x}(t), \gamma) \tag{6.1}$$

where, $\mathbf{x}$ is a vector of n_v variables, γ is a set of n_p parameters in the model and $\mathbf{f}(\mathbf{x}(t))$ is a vector of functions. The model is a mathematical representation which describes approximately some physical process. Associated with this process is a set of

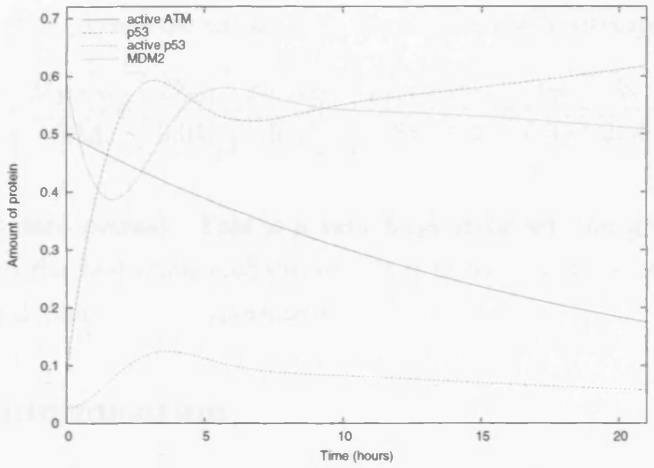


Figure 6.1: The model solution used to construct the pseudo data. The initial values $[ATM] = 0.5$, $[p53] = 0.0797$, $[Ap53] = 0.0271$ and $[MDM2] = 0.587$ and the test parameter values (Table 6.1) are used.

data gathered from experiments. Each variable in the model is measured at a number of different time points, $t = t_i$ where $i = 1 \ldots n_t$ (n_t is the total number of time points). The vector of data at $t = t_i$ is defined as follows,

$$\hat{\mathbf{x}}(t_i) = \hat{\mathbf{x}}_i.$$

The problem is to find the parameters, γ, such that the model agrees with the data in the best possible way. In the maximum likelihood realisation this is normally defined by the minimum of the least-squares value,

$$l(\gamma) = \sum_{i}^{n_t} \left\| \frac{\hat{\mathbf{x}}_i - \mathbf{x}(t_i)}{\boldsymbol{\sigma}_i} \right\|^2, \tag{6.2}$$

where $\boldsymbol{\sigma}_i$ is a vector of n_v values that are the standard deviation of the error distribution at t_i for each component of the model. This is guaranteed to provide the maximum likelihood estimation providing it is assumed that the measurement errors are Gaussian and independent (Press *et al.*, 2002) (see appendix B.1 for proof). When the error is not known it is assumed that the errors are identically distributed (every value of σ_i is set to one).

6.1.2 Example system

In this chapter various parameter estimation techniques will be applied to model 1 (equation 4.1) to assess the most appropriate techniques to use on the proposed p53 models. To this end, a data set is constructed by sampling from a run of the model (Figure 6.1). The parameter estimation routines will attempt to reproduce the parameter values used in the run (Table 6.1). In this chapter the data set size will generally be fixed at 1000

Table 6.1: The parameter values that the routines will attempt to recover.

D_{ATM}	D_{MDM2}	D_{p53}	p_{MDM2}	p_{p53}	k_1	k_2	k_3	k_4
0.05	0.18	0.041	0.01	0.52	1.42	0.39	2.50	0.75

time points (4000 data points). This is a very large data set and gives the parameter estimation routines the best chance of success. It is unrealistic to expect this amount of data from biological time course experiments.

6.2 Local minimisation

6.2.1 Rejection of some groups of parameter estimation routines

Many multidimensional optimisation algorithms have been proposed and here only a limited subset will be examined. General linear least squares is an effective technique that uses linear algebra to find the exact solution in one step (Press *et al.*, 2002). This method relies on the model solution being a linear combination of basis functions. This is generally not the case for differential equations and so this approach is unsuitable here. There is also a group of optimisation methods that use gradients in parameter space to find the minimum, the most popular of which is the Levenberg-Marquardt method (Marquardt, 1963; Stortelder, 1996). For the problem presented here this would require numeric differentiation which adds considerably to the complexity of the implementation, especially if multiple models are being examined. Therefore, these methods were avoided.

6.2.2 Nelder-Mead parameter estimation algorithm

The Nelder-Mead or downhill simplex method is based on a simple concept that can be easily implemented and understood. It is an algorithm that finds the local minimum of a function by manipulating a simple geometric object called a simplex (see appendix B.2 for more information) (Nelder and Mead, 1965). This algorithm only has a few factors that need to be set by the user (how the initial simplex is constructed, the stopping condition, and the function to be minimised) and can be easily adapted to diversify its usage especially in the context of global minimisation methods. Also, it is easy to incorporate a minimisation function that requires the integration of an ODE model. The main disadvantage of this approach is that it can require a large number of function calls.

The Nelder-Mead parameter estimation algorithm was implemented in C++. The initial simplex that seeds the algorithm was constructed around the best guess of the parameters ($\mathbf{P}_0$); each vertex is placed a fixed length, λ, from $\mathbf{P}_0$ along each dimension of the parameter space.

$$\mathbf{P}_i = \mathbf{P}_0 + \lambda \hat{\mathbf{e}}_i$$

where $\hat{\mathbf{e}}_i$ is the unit vector along the ith dimension. This is the construction method

Table 6.2: The points in parameter space used as the initial "best guess" point of the simplex.

Label	D_{ATM}	D_{MDM2}	D_{p53}	p_{MDM2}	p_{p53}	k_1	k_2	k_3	k_4
A	0.05	0.18	0.041	0.01	0.52	1.42	0.39	2.50	0.75
B	1	1	1	1	1	1	1	1	1
C	9	13	9	23	60	33	2	53	10
D	5	10	1	8.1	0.02	6.7	2.1	0.3	7.7

Table 6.3: The parameter estimates obtained from testing the Nelder-Mead optimisation method. Four different initial points were tested (see Table 6.2). it = number of iterations before convergence. LSQ = minimum least squares value.

Point	D_{ATM}	D_{MDM2}	D_{p53}	p_{MDM2}	p_{p53}	k_1	k_2	k_3	k_4	It	LSQ
A	0.05	0.18	0.041	0.01	0.52	1.42	0.39	2.50	0.75	1379	1.16×10^{-8}
B	0.0502	0.844	−0.637	0.353	0.374	2.27	0.282	3.02	0.995	3697	0.363
C	0.0509	26.5	21.0	13.3	27.6	43.5	21.9	48.2	21.6	604	7.16
D	0.0498	−14.8	42.1	4.26	42.6	2.67	−61.1	0.3	−21.3	1690	5.27

that is suggested by Press *et al.* (2002). The algorithm is stopped when the accuracy required, η, is larger than the fractional range of the simplex,

$$\text{Stopping measure} = 2.0 \times \frac{|l_h - l_l|}{|l_h| + |l_l|},$$

where l_h is the highest function value and l_l is the lowest function value held by a vertex. For each error function value calculated the model needs to be integrated so the least squares value can be evaluated. Therefore, a large proportion of time is spent integrating the model and so a fast and accurate integrator is required. Here a Runge-Kutta algorithm with adaptive step size control is used (see appendix B.5).

6.2.3 Nelder-Mead algorithm experiments

The Nelder-Mead algorithm was applied to a series of starting conditions (see Table 6.2). In all experiments the length scale, λ, and the accuracy required, η, was kept constant. λ was set to 10, a similar order of magnitude as the true parameter values and η was set to 10^{-10} (Press *et al.* (2002) suggest that η should be set at the machine precision or slightly larger, in this case the machine precision was 10^{-14}).

When the initial point is set to the minimum (point A), the simplex collapses down around this point (Table 6.3). This shows that the downhill simplex method does work in this case providing the initial point is close to the true solution. For all other initial points, the true solution is not reached indicating that there is a large number of local minima. Multiple local minima are common when applying parameter estimation to ODE models (Esposito and Floudas, 2000), especially when the model is non-linear and there are a large number of parameters. Apart from initial point A, point B produces the best result and starts closest to the true solution, suggesting that the initial distance

Table 6.4: The parameter estimates obtained from testing the Powell's method using a length scale of 10 and an accuracy of 10^{-10}. Four different initial points were tested (see Table 6.2). it = Number of iterations before convergence. LSQ = minimum least squares value.

Point	D_{ATM}	D_{MDM2}	D_{p53}	p_{MDM2}	p_{p53}	k_1	k_2	k_3	k_4	it	LSQ
A	0.05	0.18	0.041	0.01	0.52	1.42	0.39	2.50	0.75	27	1.41×10^{-8}
B	0.0502	0.272	0.129	0.0471	0.504	1.21	0.386	2.69	0.81	1530	0.0407
C	0.0499	3.08	5.87	1.62	0.545	−9.66	0.407	−0.383	−0.118	1458	2.86
D	0.0502	13.6	−1.47	6.66	0.312	3.66	0.235	22.9	9.15	495	1.44

from the solution is important. This method will fail unless a good initial estimate is available, but this is unlikely for biological systems such as the p53 network. In other tests not shown here it was confirmed that this result does not depend on the details of the implementation of the Nelder-Mead algorithm[1] or the size of the data set used .

Surprisingly, for all initial points the parameter estimate for D_{ATM} is within 0.001 of the true value. This suggests that it is particularly important to get D_{ATM} accurate, this probably occurs because ATM is the driver of the system, so if D_{ATM} is off it affects not only the solution for the amount of ATM but all other components as well producing a high error function value.

It became clear whilst performing these experiments that a large proportion of parameter sets produced results where at least one component of the model approached infinity. These parameter sets make the model stiff which causes the numerical integration to become inefficient; this is a common problem for parameter estimation (Tjoa and Biegler, 1991). This means that the parameter space has very steep hills which could cause problems for any minimisation algorithm that is implemented.

6.2.4 Direction set (Powell's) method

To confirm that the parameter space has many local minima another multi-dimensional minimisation method was implemented and tested: Powell's method (see appendix B.3) (Acton, 1990). Powell's method performs line minimisations along a series of directions; the set of directions are updated repeatedly so that directions that cause the maximum improvement are used. This method was run on the problem with each of the four initial points (Table 6.2). The results confirm that there are multiple local minima in the parameter space (Table 6.4), with the local minima found being different from those found by the downhill simplex method. As expected point A, the actual minimum, produces accurate results. Surprisingly, the results were reasonably good for point B. Apart from initial point A, the final least squares value was an improvement over that found with the Nelder-Mead approach, this may be because the accuracy value is more stringent in Powell's method.

[1]This includes how the initial simplex is constructed, the value of λ and the accuracy required.

6.3 A look at the parameter space

It is difficult to get a complete picture of the parameter space due to the large number of parameters and hence the high dimensionality of the space. Here one dimensional and two dimensional cuts through the parameter space are examined. Initially, cuts were made through the global minimum (see Figure 6.2). There is a general smooth increase in the least squares value as the parameters move away from the global minimum. There are no other minima suggesting that if all the other parameters are correct the global minima is easily found. Generally the least squares value increases more sharply when the point is moved negatively away from the minimum, especially when the parameter value becomes negative. This is reasonable as biologically all parameters should hold positive values. D_{ATM} in particular has very steep slopes around its true values.

A large range of behaviour was found when two parameters were varied around the global minimum (Figure 6.3). Generally when the parameters take negative values the least square value rises sharply (for example Figures 6.3(a)–(c)), this behaviour is particularly apparent when both parameter values are negative. The approach to the global

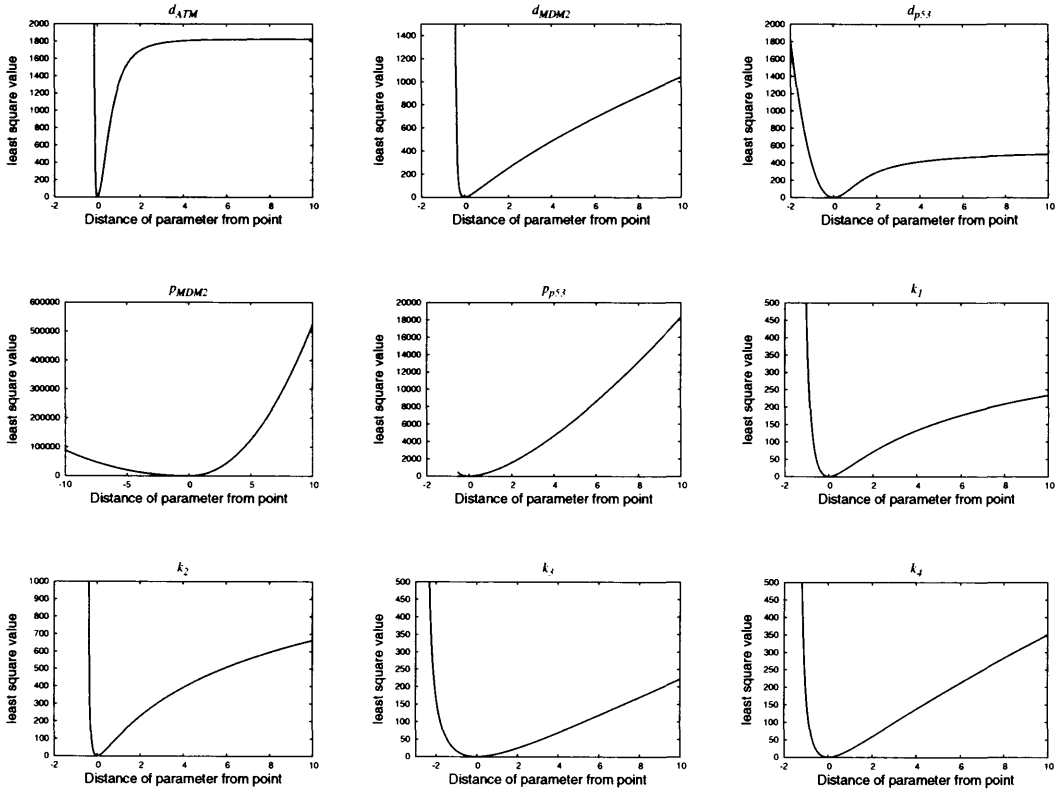


Figure 6.2: Charts to show how the least squares metric varies around the global minimum for each parameter. Each parameter was varied between a distance of -10 to a distance of +10 away from the reference point whilst the other parameters were kept fixed. Points that have gone out of bounds ($\geq 10^{100}$) have been excluded.

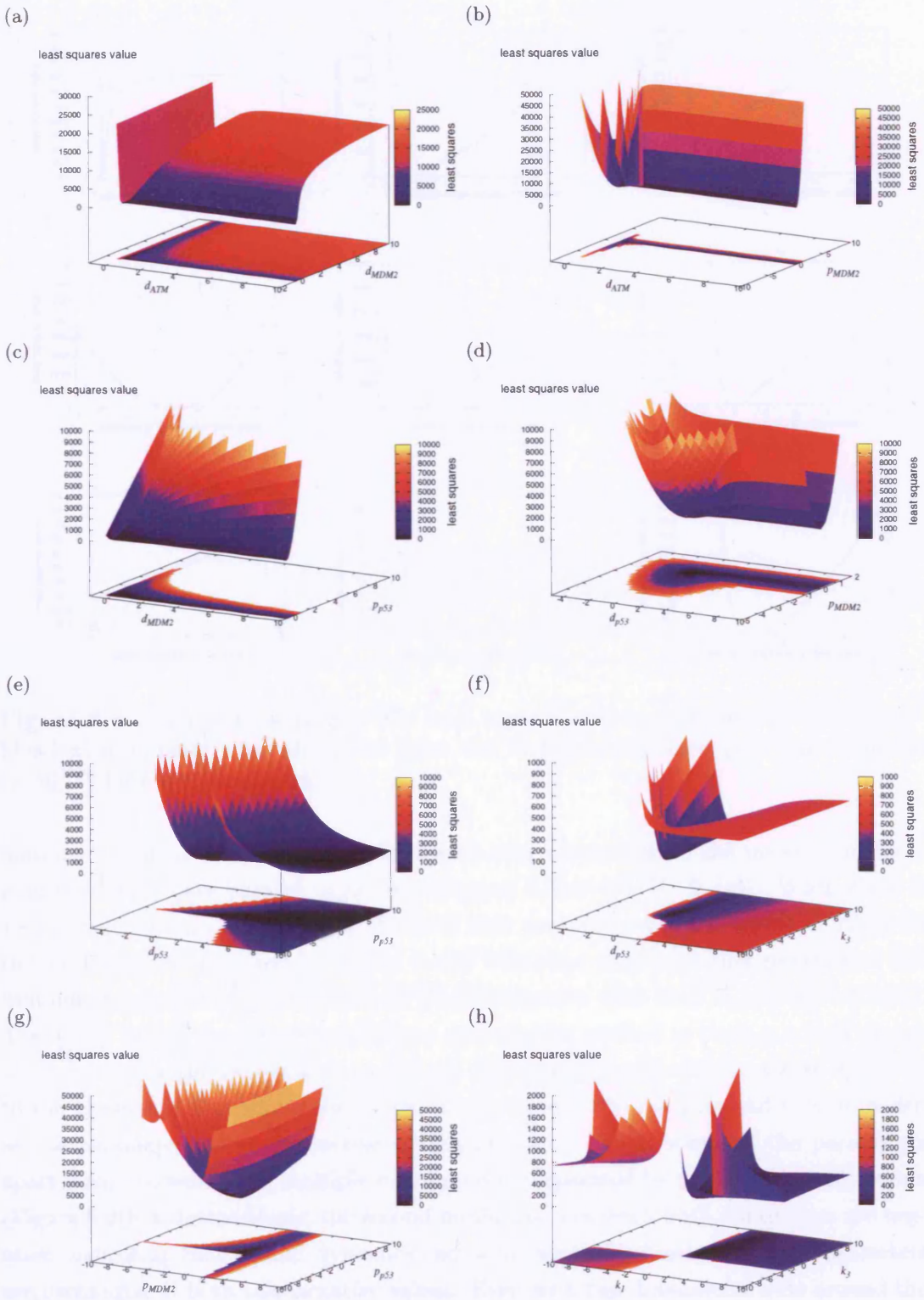


Figure 6.3: A few examples of how the least squares metric varies around the global minimum for two parameters. The axes show the difference between the estimated values and the true values.

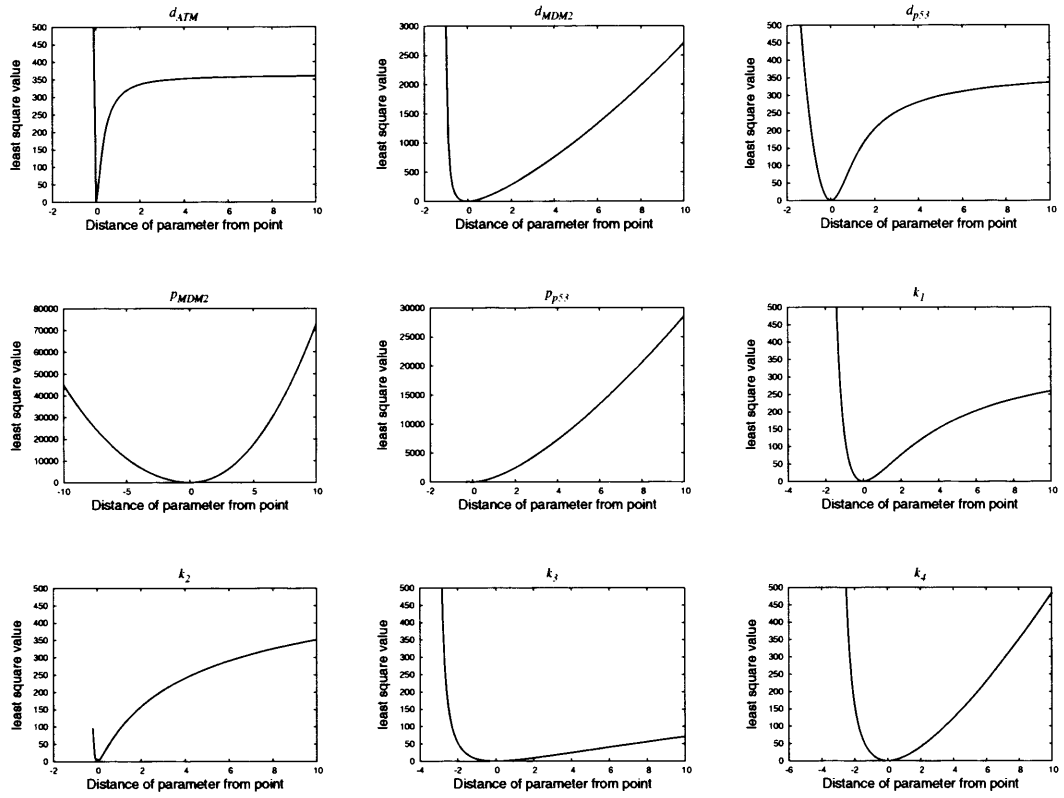


Figure 6.4: Charts to show how the least squares metric varies around the Nelder-Mead solution point when the initial point was B. Points that have gone out of bounds ($\geq 10^{100}$) have been excluded.

minimum tends to be along thin valleys with sharp boundaries. The majority of these long thin valleys are parallel to an axis (Figures 6.3(a)–(d), (f) & (g)). When there is a thin valley along only one axis, it shows that one parameter has a greater effect on the least squares value and hence the model behaviour than the other parameter. For example, p_{MDM2} has a greater effect on the least squares value than D_{p53} (Figure 6.3(d)). These thin valleys cause problems for any optimisation method as there is a high chance of stepping over the valleys and once in the valley there would only be a slow approach to the minimum. Sometimes the valleys are not parallel to the axes and tend to widen as the parameter values are increased (Figure 6.3(e)). Even when all the parameters apart from two were fixed multiple minima occur, separated by an out-of-bounds region (Figure 6.3(h)). Interestingly, the second minima occurs when both parameters are negative, indicating that similar dynamics can occur when the function of two parameters are exchanged, if both take negative values. Even with two dimensional cuts around the global minimum there is a considerable amount of complexity in the parameter space, even when the amount of data is large.

One dimensional cuts around the solutions found by the Nelder-Mead algorithm (see Table 6.3) indicate that a minimum has been found (for example Figure 6.4). This

suggests that the algorithm is working correctly and the parameter space is full of local minima. For all parameters apart from p_{MDM2} the least squares value raises more rapidly when moving negatively away from the point and many move "out of bounds". As the parameter value increases, the least squares continues to increases but for some parameters such as D_{ATM}, D_{p53}, k_1 and k_2 the least squares value seems to saturate. This indicates that other effects are regulating that mechanism.

Along a line between two of the solutions found by the Nelder-Mead algorithm (Figure 6.5(a)), again the solutions appear to be local minima. The two minima are very different, the solution from initial point C is in a shallow wide valley, whilst the solution from initial point D is in a narrow valley. There is also an area of "out of bounds" between the two points. These areas seem to be prevalent in the parameter space making it difficult for any optimisation algorithm to function. As discussed above it appears that all the solutions found by the Nelder-Mead algorithm are local minima, but it was found that for initial point B that this is not true; there is no barrier along the direction between the solution and the global minimum (see Figure 6.5(b)). The Nelder-Mead method samples possible directions and so it is possible that a beneficial route will be missed if access to it is narrow. One possible way to minimise this effect is to slow down the rate of contraction so that more directions are sampled before convergence.

In this section some of the complexity of the parameter space has been revealed. It shows the difficulty that any minimisation routine has when confronted with this optimisation problem. If the landscape is this complex for two free dimensions it will be many times more complicated when all of the parameters are allowed to vary.

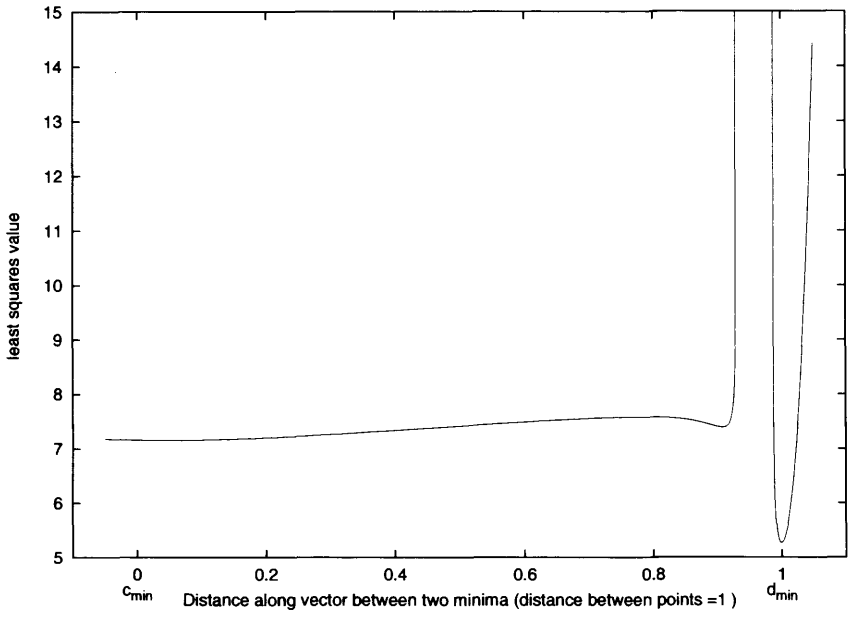
6.4 Simple approaches to global minimisation

Due to the abundance of minima advanced methods that have the ability to find the global minimum need to be used. In this section, two simple methods will be looked at that are adaptations of the basic downhill simplex method. In the next section a more complex approach will be considered.

6.4.1 Repeatedly reinitialising the simplex once a minimum is reached

This method is a simple continuation of the downhill simplex method. When a local minimum is reached the downhill simplex method is restarted with the minimum point used as the initial "best guess" point. The initial simplex is reinitialised and the downhill simplex method is run again. This is repeated until no further improvement in the minimum LSQ value is made. The idea is that once the simplex gets stuck in the local minimum that the reinitialisation of the original simplex will increase its spread wide enough to remove it from the local minimum. If the simplex converges to the global minimum then the simplex with collapse back around the original point. This algorithm was run on initial points, B, C and D (see Table 6.2).

(a)



(b)

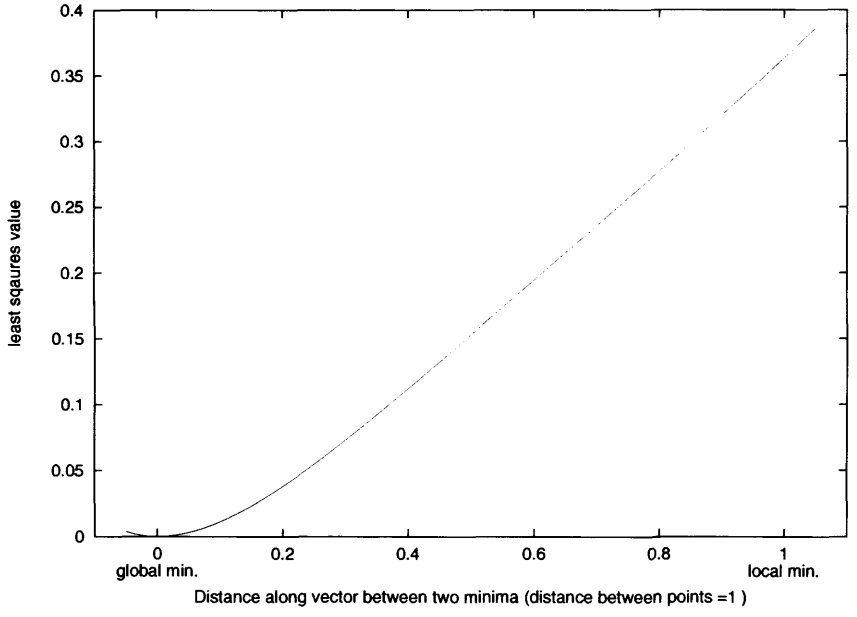


Figure 6.5: A plot of how the least squares metric varies along a line joining a) c_{min} and d_{min} (where c_{min} is the solution found when initial point C was used in the Nelder-Mead method and d_{min} is the solution found when initial point D was used (see Table 6.3)) and b) the global minimum and the local minimum found by using initial point B.

Table 6.5: A summary of the results obtained from testing the restart downhill simplex method on a range of points (see Table 6.2). The length scale, λ was set at 10 and the accuracy required to 10^{-10}. LSQ is the least squares measure and R is the number of re-initialisations before convergence.

Point	D_{ATM}	D_{MDM2}	D_{p53}	p_{MDM2}	p_{p53}	k_1	k_2	k_3	k_4	R	LSQ
B	0.05	0.18	0.041	0.01	0.52	1.42	0.39	2.50	0.75	2	1.16×10^{-8}
C	0.0501	254	−232	133	0.284	445	0.211	4.15	1.53	3	2.26
D	0.0502	122	28.9	63.8	0.292	−54.6	0.216	−15.0	−5.64	6	2.40

Initial point B was the only initial point that caused the algorithm to converge to the correct parameter values (Table 6.5). This is probably because the minimum after one Nelder-Mead run is not a true local minima (see above). Despite the improvement over the standard Nelder-Mead method in the least squares values for all initial points, point B and C have some parameter estimates further from their true values (D_{MDM2}, p_{MDM2}, k_1). The simplex restart minimising routine is too dependent on initial conditions to be of practical use. A possible improvement would be change the stopping condition so that it was more adaptive to the least squares values, for example the initial size of the simplex could be made proportional to the least squares.

6.4.2 Nelder-Mead method with momentum

Like all global minimisation techniques this method forces the algorithm to make steps that are not locally desirable. This approach adds momentum to the system, adding a proportion of the step made in the last iteration to the current step (Gershenfeld, 1999). Like a ball the simplex will continue to roll after it reaches a local minimum and hopefully roll out of the basin of attraction. The proportion of the last step applied must be enough to roll the simplex out of local minimum but small enough so it does not roll off to infinity. If $\mathbf{P}(t)$ is the simplex at iteration t then:

$$\mathbf{P}(t) = \mathbf{P}(t-1) + \Delta\mathbf{P} + \alpha[\mathbf{P}(t-1) - \mathbf{P}(t-2)]$$

where $\Delta\mathbf{P}$ is the step that would be introduced by the downhill simplex method and α is the proportion of the last step that is carried over to the current step.

This method was tested with a number of different α values using initial points B and D. When the momentum factor is too large the least squares value gets so large that it goes out of bounds (Tables 6.6 and 6.7). For smaller αs there are mixed results and no obvious correlation between α and the least squares value. There are sometimes improvements in the least squares value but at other times the least squares value gets considerably worse. For initial point B there was no improvement. At no tested value of α was the least squares value small enough for the parameters to be usable. This method is far too sensitive to the momentum parameter α. A possible improvement could be to vary α according to some factor, for example the least squares value.

Table 6.6: A summary of the results obtained from testing the downhill simplex with momentum for initial point B (see Table 6.2).

Momentum factor (α)	Iteration Number	Minimum LSQ value
0	3697	0.363
0.01	1489	0.535
0.1	1924	1.39
0.2	1482	4.92
0.5	1099	5.76
0.9	N/A	∞

Table 6.7: A summary of the results obtained from testing the downhill simplex with momentum for initial point D (see Table 6.2).

Momentum factor (α)	Iteration Number	Minimum LSQ value
0	1690	5.27
0.01	874	4.60
0.1	1323	1.93
0.2	2535	2.43
0.5	975	117
0.9	N/A	∞

6.4.3 The simple global minimisation methods combined

The momentum and the restart method were combined and tested using a range of momentum factors and initial points C and D. It is a set of mixed results (Tables 6.8 and 6.9), at certain momentum factors the global minimum was found but for other αs the result is worse than when there is no momentum. The global minimum was only found for initial point D which suggests that the performance of this method depends on starting conditions. There is no clear relationship between the value of α and the minimum least squares values, but a small non-zero α tends to give the best results

In some cases a large number of re-initialisation were required before convergence. This reveals that the parameter space contains a very high number of local minima and provides an explanation as to why the parameter estimation routine is not consistent in its ability to find the global minimum. The combination of methods worked well in certain situations and probably could be further improved by intelligently resizing the initial simplex if it converges to a point with a large error function value. When applied to real experiment data a large number of momentum factors would have to be tried before one could be certain that the global minimum had been reached. On smaller data sets and larger models though this combined method becomes very unreliable with it unable to find the global minimum. This method cannot be practically used and so more sophisticated techniques need to be examined.

Table 6.8: The results obtained from testing the downhill simplex with momentum and restart method using initial point C (see Table 6.2), $\eta = 10^{-10}$ and $\lambda = 10$.

Momentum factor (α)	No. of Restarts	Minimum LSQ value
0	3	2.26
0.01	3	2.09
0.1	6	5.83
0.2	49	2.65
0.5	29	10.7

Table 6.9: The results obtained from testing the downhill simplex with momentum and restart method using initial point D (see Table 6.2), $\eta = 10^{-10}$ and $\lambda = 10$.

Momentum factor (α)	No. of Restarts	Minimum LSQ value
0	6	2.40
0.01	4	1.17×10^{-8}
0.1	115	1.17×10^{-8}
0.2	1507	0.00233
0.5	2	7.69

6.5 Simulated annealing

6.5.1 Introduction

The simple global minimisation approaches have failed to reliably produce satisfactory results. Therefore, in this section a more sophisticated global minimisation method called simulated annealing will be examined. Simulated annealing is based upon the thermodynamical process called annealing (Gershenfeld, 1999). The system is assigned a temperature, T. The higher the temperature, the more likely the system will take a step to a position with a worse error function value. Initially the temperature starts at a high value and so most steps are accepted allowing the space to be thoroughly sampled. The temperature is then gradually decreased until only steps that improve the position are accepted. The idea is that the system will naturally find its way to the basin of attraction of the global minimum. It has been found to be suitable for large scale optimisation problems, especially those where the global minimum is hidden among many local minima (Press *et al.*, 2002). More details can be found in appendix B.4. There are many other global minimisation algorithms that could be used, such as genetic algorithms (Gershenfeld, 1999) and Markov Chain Monte Carlo (Gilks *et al.*, 1996). All of these take a similar approach; they intelligently sample the solution space so that the global minimum can be found.

The key elements of a simulated annealing implementation are the cost measure, the method to propose the next step and the scheme to cool the temperature. For parame-

ter estimation the likelihood and hence the least squares measure is an appropriate cost measure. A popular approach used to propose the next step is to use the Nelder-Mead method. This was first introduced by Press *et al.* (2002) and has been implemented numerous times to good effect (Torres *et al.*, 1997; Kvasnicka and Pospichal, 1997; Cardoso *et al.*, 1996). It is similar to the Nelder-Mead routine apart from that at the beginning of each iteration a thermal fluctuation, $-T \ln p$ (where p is a random number taken from a uniform distribution between 0 and 1), is added to the real function value of each point in the simplex. Also for any point that is proposed by the simplex routine a thermal fluctuation is taken from the function value of the proposed point i.e. $f = f + T \ln p$ where f is the proposed function value. At a high temperature a proposed point is likely to be accepted even though it has a higher function value than points in the simplex. As the temperature approaches zero this method reduces to the standard downhill simplex method. An alternative method to propose points would be to take a random point from a Gaussian distribution with the mean set at the current point and the standard deviation set at the length scale. This was found to be ineffective compared with the Nelder-Mead proposed step and makes little sense since the chance of guessing a good direction in a high-dimension parameter space is slim (Gershenfeld, 1999).

The cooling scheme used is one recommended by Press *et al.* (2002): $T = T_0(1-k/K)^4$ where T_0 is the initial temperature, k is the total number of moves so far and K is the estimated number of moves required. K is an important factor; if the the system is cooled too fast the solution will be a local minimum and if it is cooled too slowly there is not only the waste of computer resources but the potential that the system will move irrevocably out of computational bounds. T_0 needs to be high enough so that initially at least 50% of proposed states are accepted. At $T = 0$ the simplex is shifted to the point with the lowest least squares value encountered.

6.5.2 Experiments

To examine the effectiveness of the simulated annealing method a number of different experiments were performed, with different initial points, initial temperatures (T_0) and total number of moves (K). Repeats were also performed to check how robust the method was. The initial simplex configuration should not affect the algorithm too much as the initial temperature should be high enough to allow most moves so effectively randomising the initial simplex. The initial temperatures of 10, 100 and 1000 were used and have a corresponding initial acceptance percentage of 52%, 68% and 71% (this is the percentage of proposed steps that are accepted based on the 1000 steps proposed after the first 100 steps). The total number of steps were chosen for their feasibility; 10^6 steps can take up to a couple of days so it was chosen to mainly use that number of steps and occasionally use 10^7 steps.

The results are generally good when initial point B is used, with the majority ending at the global minimum with a least squares value of 6.36×10^{-4} (see Table 6.10 for a

Table 6.10: The results from using simulated annealing with downhill simplex parameter estimation. A range of initial temperatures, estimated counts and initial points were used.

Initial point	Starting Temperature (T_0)	Estimated count (K)	Number of iterations	Least squares score
B	10	10^6	1000000	6.36×10^{-4}
B	10	10^6	1000000	6.36×10^{-4}
B	10	10^6	1000000	6.36×10^{-4}
B	10	10^6	1000000	6.36×10^{-4}
B	10	10^6	1000000	0.288
B	10	10^7	10000000	6.36×10^{-4}
B	10	10^7	10000000	0.237
B	100	10^6	1000000	6.36×10^{-4}
B	100	10^6	1000000	6.11
B	100	10^6	1000002	16.4
B	100	10^7	10000000	11.4
B	1000	10^6	1000000	110
C	10	10^6	1000001	5.22
C	10	10^6	1000001	5.39
C	10	10^6	1000000	5.48
C	10	10^6	1000000	5.77
C	10	10^6	1000000	5.98
C	10	10^7	10000000	5.37
D	10	10^6	1000000	4.28
D	10	10^6	1000000	5.50
D	10	10^6	1000001	5.67
D	10	10^6	1000002	5.71
D	10	10^6	1000000	12.4

summary and Table D.1 in appendix D.1 for the full results). The global minimum does not have a least squares as low as previously because the accuracy of the embedded Runge-Kutta was reduced to improve the speed of the algorithm. When $T_0 = 10$, all the runs reach the minimum apart from two (the corresponding parameter values are still reasonably good). When the initial temperature is higher then 10 the algorithm is more likely to give poor results. If different initial points are used the global minimum is not found and the parameters are very poor, even if the estimated count is increased. The parameter estimates generally do not show any relation to the true values apart from D_{ATM} which is within 10% of the true value in 10 out of 11 runs. These results indicate that initial point B must be a special case, as mentioned previously. The actual number of iterations for all runs is very close to the estimated count, which suggests that the system is at a local minimum before the temperature reaches zero.

These results suggest that simulated annealing is not working satisfactorily. A possible reason for this is that the initial temperature or the estimated count is not large enough, but the estimated count is as large as it can be for the algorithm to be feasible

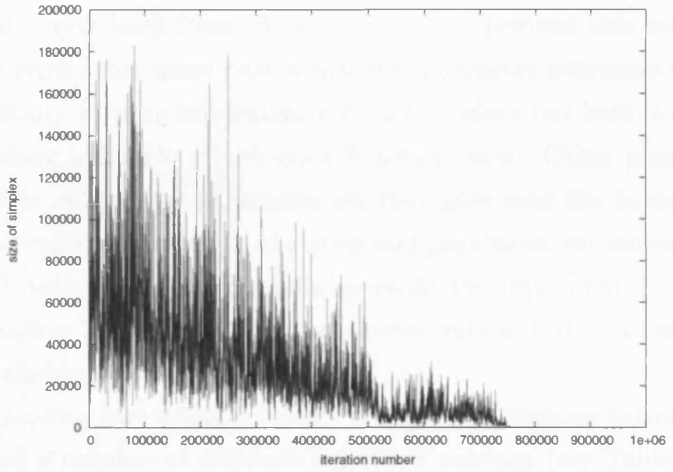


Figure 6.6: A chart to show how the size of the simplex varies over the course of a simplex simulated annealing run. The size of the simplex is measured by the averaged squared distance between the centroid of the simplex and its vertices. The run had started at initial point C, had an initial temperature of 10 and an estimated count of 10^6.

and the results suggest that increasing the temperature increases the likelihood of getting a poor result. Another possibility is that the shape of the cooling scheme is inappropriate; only one of the three schemes suggested by Press *et al.* have been tried and there are many other schemes such as the saw tooth scheme suggested by Torres *et al.* (1997). This is tinkering with the detail and there is no guarantee that a cooling scheme that works well for one model will work well for others.

Another problem is due to large parts of parameter space having out of bounds error function values. During the initial stages of simulated annealing the temperature is high and the majority of transformations are accepted causing the simplex to rapidly increase to a very large size (Figure 6.6) and pushing the centroid of the simplex far away from the global minimum. As the temperature decreases, the simplex finds it difficult to come together in an optimal way due to the large proportions of parameter space that produce out of bound values. This makes it much more likely that the simplex will end up in a local minimum far from the global minimum as has been seen in the experiments performed here.

Simulated annealing does not reliably produce the correct parameter estimates even at large amounts of data. This result was confirmed by applying the algorithm to a number of different models, data set sizes and error functions. Various minor alterations were made to the algorithm, such as varying the number of times the simplex was shifted to the best parameter set. In none of these configurations were the results satisfactory.

6.5.3 Adding boundaries

One problem with simulated annealing is that when the temperature is high, the simplex has the tendency of becoming very large and reaching large parameter values that it

cannot successfully move back from. A possible way to prevent this would be to include boundaries in the parameter space past which the parameter estimates would become invalid. There are many ways to implement such a boundary but here, a set of parameters beyond the boundary will have a high error function value. Other possible implementations are absorption, where any parameter set that goes over the boundary is moved to the boundary, and reflection where if the proposed parameter set moves over the boundary it is reflected back. Realistic boundaries could be implemented if there was some additional information known about the parameter values, but it is rare for this kind of information is available.

Simulated annealing was applied to the system with various boundaries on the parameter values and a number of different algorithm settings (see Table 6.11). When the parameter values cannot take negative values, the parameters are reasonably close to the true values if the initial point is B but are distant if the initial point is D. This boundary does not appear to have produced an improvement. When the parameter values are restricted to a tight region around the true values (between 0 and 3) the results are generally good even though in most cases the global minimum has not been found. In some cases certain parameter values appear to get stuck at a boundary, particularly for parameters p_{MDM2} and k_3 which have the smallest and largest true value respectively. This probably happens because the boundary was implemented as a sharp boundary, smoothing the boundary might alleviate this problem. When a larger range of parameter values are accepted and initial point D is used, a similar set of results is obtained, with one run in particular producing very good values. This shows that restricting the range of parameter values can improve the chance of finding reasonable values.

As there was some success simulated annealing with boundaries was applied to a smaller data set of 40 time points. The global minimum was sometimes found and sometimes not, but the final least squares value was always reasonably good and the results are significantly better than when there were no bounds (Table 6.12). Even when not at the global minimum, at least some of the parameter estimates are close to their true values and at least one of the estimates is at the boundary. It is concerning that the least squares value can be very low, for example 8.93×10^{-5}, and still have one or two parameter estimates away from their true value; this effect is likely to be because of the lower amount of data so that more model solutions can fit the data equally well. When simulated annealing was applied to a 20 time point data set the results were about as reliable as the 40 time point set with two out of ten reaching the global minimum when the boundaries were 0 and 10, and four out of fourteen reaching the global minimum when the boundaries were 0 and 100. The global minimum of 1.00×10^{-5} and the corresponding parameters were a bit further away from the true values than the 40 time point data set results (see Table 6.13).

Table 6.11: A summary of the results obtained from simulated annealing on a data set with 1000 time points. Two initial points were used (see Table 6.2), $T_0 = 10$ and the number of estimate steps was set to 10^6 or 10^7 (K). LSQ is the resulting least squares value.

Initial Point	Boundaries	K	LSQ	D_{ATM}	D_{MDM2}	D_{p53}	p_{MDM2}	p_{p53}	k_1	k_2	k_3	k_4
True	–	–		0.05	0.18	0.041	0.01	0.52	1.42	0.39	2.50	0.75
B	0	10^7	0.00283	0.0500	0.162	0.0656	4.50×10^{-58}	0.532	1.41	0.399	2.50	0.748
B	0, 3	10^6	0.000636	0.0500	0.200	0.0445	0.0187	0.521	1.42	0.391	2.53	0.755
B	0, 3	10^6	0.000786	0.0500	0.193	0.0419	0.0159	0.522	1.42	0.392	2.52	0.752
B	0, 3	10^6	0.0535	0.0498	0.185	0.199	2.69×10^{-18}	0.557	1.24	0.401	3	0.860
B	0, 3	10^7	0.00283	0.0500	0.162	0.0656	9.74×10^{-76}	0.532	1.41	0.399	2.50	0.748
B	0, 3	10^7	0.0305	0.0498	0.271	0.0826	0.0400	0.504	1.30	0.367	3	0.848
D	0	10^6	5.52	0.0513	1490	1110	674	708	64.6	730	2.11×10^{-10}	1.93×10^{-8}
D	0	10^6	5.68	0.0501	1630	1390	2010	878	0.000127	625	1560	2.82×10^{-9}
D	0	10^6	6.02	0.0518	1580	945	436	580	116	749	2.06	3.29×10^{-10}
D	0, 10	10^6	0.000636	0.0500	0.200	0.0445	0.0187	0.521	1.42	0.391	2.53	0.755
D	0, 10	10^6	0.00283	0.0500	0.162	0.0656	1.02×10^{-17}	0.532	1.41	0.399	2.50	0.748
D	0, 10	10^6	2.52	0.0508	10	5.17×10^{-15}	5.84	0.705	2.06	0.564	10	10

Table 6.12: A summary of the results obtained from simulated annealing on a data set with 40 time points and using initial point B (see Table 6.2). The number of steps was set to 10^6 and $T_0 = 1000$. LSQ is the resulting least squares value.

Boundaries	D_{ATM}	D_{MDM2}	D_{p53}	p_{MDM2}	p_{p53}	k_1	k_2	k_3	k_4	LSQ
True	0.05	0.18	0.041	0.01	0.52	1.42	0.39	2.50	0.75	–
0, 3	0.05	0.199	0.0466	0.0187	0.522	1.42	0.392	2.53	0.755	2.44×10^{-5}
0, 3	0.0499	0.196	0.052	0.016	0.5233	1.41	0.392	2.54	0.752	5.9×10^{-5}
0, 3	0.05	0.162	0.068	4.61×10^{-13}	0.533	1.41	0.4	2.51	0.747	1.12×10^{-4}
0, 3	0.05	0.205	3.88×10^{23}	0.0241	0.517	1.49	0.388	2.49	0.753	6.04×10^{-4}
0, 3	0.0503	2.22	4.55×10^{-15}	1.17	0.516	4.48	0.405	3	1.65	0.0507
0, 3	0.0502	2.92	4.82×10^{-15}	1.52	0.578	1.67	0.447	3	1.50	0.0617
0, 10	0.05	0.199	0.0466	0.0187	0.522	1.42	0.392	2.53	0.755	2.44×10^{-5}
0, 10	0.05	0.211	3.26×10^{-22}	0.0257	0.512	1.47	0.385	2.50	0.746	8.93×10^{-5}
0, 10	0.05	0.162	0.068	3.78×10^{-13}	0.533	1.41	0.4	2.51	0.747	1.12×10^{-4}
0, 10	0.05	0.162	0.068	2.62×10^{-17}	0.533	1.41	0.4	2.51	0.747	1.12×10^{-4}
0, 100	0.05	0.199	0.0466	0.0187	0.522	1.42	0.392	2.53	0.755	2.44×10^{-5}
0, 100	0.05	0.199	0.0466	0.0187	0.522	1.42	0.392	2.53	0.755	2.44×10^{-5}
0, 100	0.05	0..211	1.34×10^{-16}	0.0257	0.512	1.47	0.385	2.5	0.746	8.93×10^{-5}
0, 100	0.05	0.211	1.34×10^{-16}	0.0257	0.512	1.47	0.385	2.5	0.746	8.93×10^{-5}
0, 100	0.0499	94.6	8.01×10^{-9}	49.6	0.52	1.49	0.39	100	49.6	0.0838

Table 6.13: The parameter values for the global minimum when there are 20 data points.

D_{ATM}	D_{MDM2}	D_{p53}	p_{MDM2}	p_{p53}	k_1	k_2	k_3	k_4
0.005	0.206	0.0456	0.0209	0.522	1.42	0.392	2.54	0.755

Table 6.14: A summary of the results obtained from simulated annealing applied to a data set with 1000 time points when two parameters are fixed. $T_0 = 10$, various estimated iteration numbers (K) and various initial points were used (see Table 6.2). LSQ is the final least squares value.

Initial point	K	D_{ATM}	D_{MDM2}	p_{p53}	k_1	k_2	k_3	k_4	LSQ
True	–	0.05	0.18	0.52	1.42	0.39	2.50	0.75	–
B	10^6	0.0500	0.181	0.524	1.43	0.394	2.50	0.747	0.00127
B	10^6	0.0508	199	398	1150	300	5540	2170	6.63
B	10^6	1920	0.0695	154	1010	789	88.1	1180	17.0
B	10^7	0.0500	0.183	0.523	1.43	0.392	2.52	0.749	0.00981
C	10^6	0.0508	199	398	1150	300	5540	2170	6.63
C	10^6	0.0508	199	398	1150	300	5540	2170	6.63
C	10^6	0.0510	205	398	1140	310	5380	2150	6.64
D	10^6	0.0500	0.181	0.524	1.43	0.394	2.50	0.747	0.00127
D	10^6	0.0463	1620	0.338	3.30	0.254	2650	−78.7	3.01
D	10^6	1010	−0.0375	0.298	0.881	−2.96	7.05	36.0	14.0

6.5.4 Fixing parameters

The optimisation algorithms could be finding it difficult to find the solution even when there are a large number of data points because of the high dimension of parameter space. In this section the effect of reducing the number of parameters will be examined. It is likely that two parameters that control the rate of the same function, for example increasing the amount of MDM2, will be more difficult to correctly estimate than having a single parameter associated with a single function. Therefore it was decided to examine the effect of fixing the basal rates that replicate the function of other components, in this case p_{MDM2} and D_{p53}.

The values of p_{MDM2} and D_{p53} were fixed at their true values and then simulated annealing was applied using a number of different initial conditions. The global minimum was sometimes reached for both initial point B and D (Table 6.14), suggesting that fixing parameters does make the problem less complex making it simpler for the method to find the solution. Generally though the parameter estimates are poor so finding the global minimum is still unreliable. These results, even though limited in scope, do suggest that if at all possible parameter values should be measured directly, as this can significantly improve the estimates for the other parameters.

6.6 Using collocation and splines

6.6.1 Introduction

The parameter estimation problem can be considered a boundary value problem; the data points are the boundaries and the aim is to find a solution that goes through these boundaries. Up until now the shooting method (Golub and Ortega, 1992) has been used, which reduces the boundary value problem to an initial value problem. A set of parameters are chosen, the model is integrated and the solution curve is compared to the boundaries. Based on this comparison the set of parameters are updated and this process is repeated. An alternative to this approach is a group of methods called projection methods (Golub and Ortega, 1992). Instead of integrating the model, the solution is approximated by a linear combination of basis functions,

$$\mathbf{x}(t) \approx \mathbf{u}(t) = \sum_{j=1}^{n_s} \mathbf{b}_j \Phi_j(t),$$

where $\Phi_j(t)$ is the jth basis function of a set of size n_s and $\mathbf{b}_j$ is a vector of n_v (the number of variables in the model) constant basis coefficients, associated with the jth basis function. There is a total of $n_s \times n_v$ basis coefficients. Easily differentiable basis functions are chosen,

$$\frac{d\mathbf{x}(t)}{dt} \approx \sum_{j=1}^{n_s} \mathbf{b}_j \frac{d\Phi_j(t)}{dt},$$

changing the problem to a set of algebraic equations. The solutions of these equations will be considerably easier and quicker than before the approximation. The basis coefficients add to the number of variables potentially making the problem more difficult to solve. This is not a new approach and has been successfully used on other optimisation problems especially in the chemical engineering field (Van den Bosch and Hellinck, 1974; Tjoa and Biegler, 1991; Tieu *et al.*, 1995; Wang, 2000; Esposito and Floudas, 2000). Here though, the details of the implementation are different.

B-splines - the chosen basis function

For this problem cubic B-splines (Figure 6.7) have been chosen as the basis functions (De Boor, 1978). The problem domain (in this case, time) is divided into a grid of equally spaced nodes, $t_1, \ldots, t_{n_g}$, with spacing h. For each node a corresponding B-spline, $B_i(t)$ is defined as follows:

$$\begin{array}{ll}
\frac{1}{4h^3}(t - t_{i-2})^3, & t_{i-2} \le t \le t_{i-1} \\
\frac{1}{4} + \frac{3}{4h}(t - t_{i-1}) + \frac{3}{4h^2}(t - t_{i-1})^2 - \frac{3}{4h^3}(t - t_{i-1})^3, & t_{i-1} \le t \le t_i \\
\frac{1}{4} + \frac{3}{4h}(t_{i+1} - t) + \frac{3}{4h^2}(t_{i+1} - t)^2 - \frac{3}{4h^3}(t_{i+1} - t)^3, & t_i \le t \le t_{i+1} \\
\frac{1}{4h^3}(t_{i+2} - t)^3, & t_{i+1} \le t \le t_{i+2}
\end{array}$$

otherwise $B_i(t) = 0$.

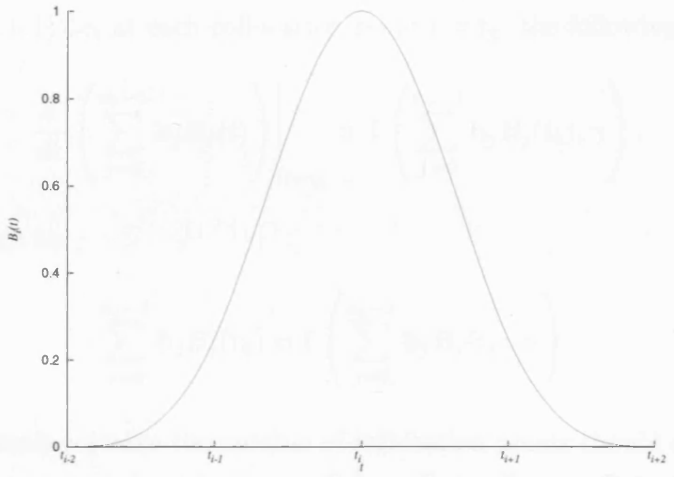


Figure 6.7: A cubic B-spline

Table 6.15: Values of B_i and B_i' at the nodes. Adapted from Golub and Ortega (1992)

	t_{i-2}	t_{i-1}	t_i	t_{i+1}	t_{i+2}
B_i	0	1/4	1	1/4	0
B_i'	0	$3/4h$	0	$-3/4h$	0

The B-spline has some useful properties: it is defined on a limited scale and on the nodes the value of B_i and the differentials of B_i are of a simple form (see Table 6.15).

A cubic spline, $c(t)$, is a piecewise cubic polynomial which has the property that it is twice continuously differentiable (Golub and Ortega, 1992) and hence smooth. ODE biological models will have solutions without discontinuities, so a cubic spline is a good choice as an approximation to the model solution. For any cubic spline defined on an equally spaced grid $t_1 < \ldots < t_{n_g}$, there are spline constants $\alpha_0, \alpha_1, \ldots, \alpha_{n_g+1}$ such that (Golub and Ortega, 1992),

$$c(t) = \sum_{i=0}^{n_g+1} \alpha_i B_i(t).$$

Two additional B-splines are needed outside of the grid (B_0 and B_{n_g+1}), this is required to satisfy the properties of the cubic spline. Hence, $n_s = n_g + 2$. The smaller h is, the greater the number of B-splines, and so the more accurate the approximation is to the real solution.

Collocation

There are a variety of ways to determine the basis coefficients of the approximate solution such as Galerkin's method and Rayleigh-Ritz but here collocation will be used (Heath, 1997). Collocation sets up a grid of n_c points (collocation points), that are not necessarily equally spaced, and requires that at these points the approximate solution satisfies the

model (Equation 6.1) i.e. at each collocation point $t = t_k$, the following is satisfied,

$$\frac{d}{dt}\left(\sum_{j=0}^{n_s-1}\mathbf{b}_j B_j(t)\right)\bigg|_{t=t_k} = \mathbf{f}\left(\sum_{j=0}^{n_s-1}\mathbf{b}_j B_j(t_k), \gamma\right),$$

which is equivalent to,

$$\sum_{j=0}^{n_s-1}\mathbf{b}_j B_j'(t_k) = \mathbf{f}\left(\sum_{j=0}^{n_s-1}\mathbf{b}_j B_j(t_k), \gamma\right).$$

For this to be exactly solvable the number of collocation points should equal the number of basis coefficients, which from here on will be called spline coefficients. If the number of collocation points is greater than the number of spline coefficients then there are more equations than variables and so the problem is said to be over-defined and there is generally not an exact solution. Therefore, the "best" set of variable values should be found, which is commonly defined as those that minimises the residual sum of squares[2]: the least squares solution. The more collocation points used (up to some limiting number) the closer the approximation, $\mathbf{u}(t)$, will be to the real solution, but for simplicity, the collocation points are set at the B-spline nodes, apart from at the extremes, B_0 and B_{n_s-1}. Therefore, $n_c = n_g = n_s - 2$.

6.6.2 Error function

Introducing the approximate solution based on B-splines increases the number of variables in the problem; n_p model parameters and the $n_s n_v$ spline coefficients. A suitable error function is,

$$\begin{matrix}\text{Error}\\ \text{Function}\end{matrix} = \sum_{i=1}^{n_t}\left\{\hat{\mathbf{x}}_i - \sum_{j=0}^{n_s-1}\mathbf{b}_j B_j(t_i)\right\}^2 + \sum_{k=1}^{n_s-2}\left\{\sum_{j=0}^{n_s-1}\mathbf{b}_j B_j'(t_k) - \mathbf{f}\left(\sum_{j=0}^{n_s-1}\mathbf{b}_j B_j(t_k), \gamma\right)\right\}^2,$$

where there are n_t time points (at t_i) and $n_s - 2$ collocation points at the nodes of the B-splines (where $t = t_k$). The requirement for this problem to be over-defined is,

$$n_t > \frac{n_p}{n_v} + 2.$$

There is no dependency on the number of spline coefficients and the 2 comes from the coefficients associated with the extreme nodes required when using B-splines.

[2]The residual of an equation is the constant that needs to be added to one side of the equation so that the equality is satisfied.

Table 6.16: The parameter estimates obtained from testing the Nelder-Mead optimisation method using the spline approximation to the model solution. The first row indicates the corresponding result when the integrator was used. It = Number of iterations before convergence, LSQ = minimum least squares value and n_s is the number of B-splines that make up the spline.

n_s	D_{ATM}	D_{MDM2}	D_{p53}	p_{MDM2}	p_{p53}	k_1	k_2	k_3	k_4	It	LSQ
N/A	0.05	0.18	0.041	0.01	0.52	1.42	0.39	2.50	0.75	1379	1.16×10^{-8}
7	0.0502	0.138	0.000423	0.0187	0.4484	1.28	0.369	1.39	0.309	32439	0.283
12	0.0503	0.175	-2.82×10^{-11}	0.0012	0.443	1.27	0.333	2.85	0.9	113098	0.00473
22	0.05	0.179	0.0209	0.00929	0.504	1.41	0.378	2.53	0.761	618884	4.27×10^{-5}
52	0.05	0.180	0.0406	0.00993	0.520	1.42	0.390	2.50	0.750	2.89×10^{7}	6.84×10^{-8}

6.6.3 Experiments

The new form of the error function was implemented and applied to the Nelder-Mead method using initial point A (see Table 6.2) and a range of spline sizes. Initial values of the spline coefficients were determined by fitting the splines to the data using a least squares fit and linear algebra.

All the parameter estimates are of the right order and the accuracy increases as n_s is increased (Table 6.16). The least squares score is always higher than when a integrator is used because the number of B-splines limits how accurate the approximation to the model solution can be; n_s would have to be very high to match the accuracy of the integrator. As n_s increases, the number of iterations it takes before convergence also increases. Even at 7 splines it takes approximately thirty times more iterations than when a integrator is used. This is because of the increase in the number of variables. Each iteration is considerably quicker than the integrator, but it still takes longer to converge; taking approximately 40 mins instead of 2 minutes when $n_s = 22$.

Simulated annealing was applied to this technique at a range of different lengths and initial temperatures. The number of B-splines was kept at 22 as this seemed a reasonable balance between accuracy and the number of extra parameters. Given that the initial point tested was B, which had good results in the standard simulated annealing, the results are disappointing as the global minimum was not found (Table 6.17 summarises the results and the full results are in Table D.2 in appendix D.1). Interestingly, the least squares are all reasonable low and the parameter estimates are the right order of magnitude. This suggests that the spline is effectively restraining the parameter estimates to values of the right order of magnitude but the simulated annealing is "cooled" too quickly to find the global minimum. The hypothesis is supported by the large number of iterations after $T = 0$ before convergence. It is impractical to increase the estimated count further. When the length of the algorithm was increased to 10^8 the resulting least squares improves by a factor of about 10 and there is a corresponding improvement in the parameters (Table D.2). When the length of the algorithm was increased to 10^9 the results were not as good. The solution splines for all the experiments are reasonably close to the true time course (Figure 6.8). The problem with this approach is that the

Table 6.17: The results from using simulated annealing with downhill simplex parameter estimation using splines. A range of initial temperatures and estimated counts were used. Initial point B was used.

Starting Temperature (T_0)	Estimated count (K)	Number of iterations	Least squares score
10	10^7	10353048	0.0289
10	10^7	10323240	0.0165
10	10^7	10413608	0.0411
10	10^7	10302822	0.0240
10	10^7	10349138	0.0215
10	10^8	100092291	0.00448
10	10^8	100207829	0.00223
10	10^8	100207386	0.00202
10	10^9	1000000000	0.0193
100	10^7	10608630	0.122
100	10^7	10431066	0.0648
100	10^7	10379737	0.0606
100	10^7	10309775	0.144
100	10^7	10376820	0.123

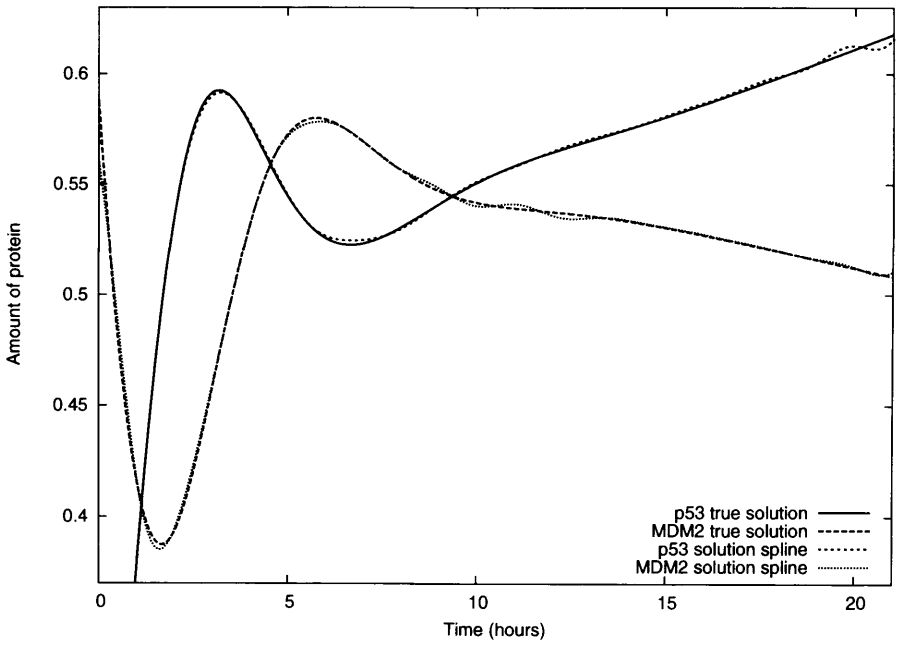


Figure 6.8: A plot to show how close the solution spline is to the true spline even when the least squares value is 0.0411. The run was started at initial point (b), had an initial temperature of 10 and an estimated count of 10^7.

number of parameters is substantially increased making the problem much more complex, without the increase in speed needed to allow simulated annealing to be run long enough in a practical amount of time. Various alterations were made to the error function, such as increasing the number of collocation points, but similar results were obtained.

6.7 Conclusion

In this chapter a range of methods have been used to try and solve the parameter estimation problem for a simple model proposed in this thesis. It is apparent that this problem is not simple; the parameter space appears to be very complex with a large number of minima and areas where the parameters produce least squares values that are out of bounds. This parameter estimation problem is non-convex with many similar local minima very close together. In this situation it is very difficult for any parameter estimation routine to be successful. Simulated annealing with a Nelder-Mead proposed step was the main global optimisation routine that was implemented. The global minimum was reached occasionally but not in a reliable manner and the rate of success depended on the starting conditions of the algorithm. It was found that by restricting the range of parameters values allowed and reducing the number of parameters the rate of success did improve but problems still persisted. To implement either of these refinements additional experiments would have to be performed. An alternative version of simulated annealing was implemented that used a spline to represent the model solution rather than integrating the model for every function call. The motivation was to constrain the parameters and speed up the algorithm. The resulting parameter values obtained were of the correct order but the global minimum was not found. The problem is that even though the spline does constrain the problem it also adds a large number of extra parameters which complicates the problem. All experiments were performed on a data set with 1000 time points. This is extremely large for a biological experiment and the performance of the methods will decrease as the number of data points is decreased. Therefore, it seems unlikely that the parameter estimation methods examined here would be of practical use when real biological data is used.

Chapter 7

A new parameter estimation method using collocation and linear algebra

7.1 Introduction

It is apparent from the previous chapter that the parameter estimation methods examined are inadequate for the p53 models. In this chapter a new parameter estimation method is developed that simplifies the problem so that more reliable and accurate parameter estimates can be found. There are two principal approximations that are made, the first of which is to use a spline as an intermediary between the model solution and the data. This has already been studied in isolation and found to be inadequate due to the large number of extra parameters that the spline requires (see section 6.6). The second approximation is to simplify the problem into a system of linear equations. Sets of linear equations have one solution that can be reliably found by algebraic techniques. This has the potential to both speed up the estimation and provide accurate results.

The technique is developed in a number of stages with each stage introducing additional refinements. In the first stage the problem is approximated as a linear system of equations. In second stage the method is adapted to include an iterative approach so that the dependence on the initial assumption is reduced. In the third stage a refinement is made to the algorithm so that more "weight" can be placed on the spline accurately representing the model solution rather than being close to the data. Finally, a number of adjustments are made to allow the algorithm to perform better when the amount of data is small. At each stage the effectiveness of the algorithm and the simplifications made are assessed using an example model and pseudo data constructed to have various quantities of error.

7.2 Setting up the problem

7.2.1 Approximating the model solution by a spline

As discussed in section 6.6 the model solution can be approximated by a spline made up of cubic B-splines,

$$\mathbf{x}(t) \approx \sum_{j=0}^{n_s-1} \mathbf{b}_j B_j(t),$$

where $\mathbf{x}(t)$ is the vector of n_v model solutions (n_v is the number of components in the model), $B_j(t)$ is the value of the jth B-spline at time t, $\mathbf{b}_j$ is the vector of n_v spline coefficients of the jth B-spline (each spline coefficient in the vector is associated with a component in the model), and n_s in the number of B-splines.

7.2.2 The system of equations

With the introduction of the spline representation of the model solution the number of variables in the problem has increased, with n_p model parameters and $n_s n_v$ spline coefficients. Ideally, the spline should satisfy the model, which is assessed through collocation

(see section 6.6.1), and also pass through the data points. There are two sets of equations:

- The data and spline equations

  At each of the n_t time points that data is measured, the following should be satisfied,

$$\frac{\hat{\mathbf{x}}_i}{\boldsymbol{\sigma}_i} = \sum_{j=0}^{n_s-1} \frac{\mathbf{b}_j B_j(t_i)}{\boldsymbol{\sigma}_i}, \tag{7.1}$$

  where $\hat{\mathbf{x}}_i$ is the vector of data measured at time $t = t_i$ and $\boldsymbol{\sigma}_i$ is a vector of n_v values that are the standard deviation of the error distribution at t_i. This is the exact algebraic version of the least squares equation (see equation 6.2) i.e. if the least squares error function was equal to zero. From here on, $\boldsymbol{\sigma}_i$ will be ignored, with the assumption that each data point (whatever the component) has the same *absolute* error.

- The model and collocation equations

  At each of the $n_c = n_s - 2$ collocation points (where $t = t_k$), the variables should satisfy,

$$\sum_{j=0}^{n_s-1} \mathbf{b}_j B_j'(t_k) = \mathbf{f}\left(\sum_{j=0}^{n_s-1} \mathbf{b}_j B_j(t_k), \gamma\right), \tag{7.2}$$

  where, γ is a set of n_p parameters in the model and $\mathbf{f}(\mathbf{x}(t))$ is a vector of model functions. Here the collocation points are taken at the nodes of the B-splines.

An exact solution could be obtained for the variables if there were exactly the correct number of equations, but only if the model perfectly represented the system and there was no error in the data. This is unrealistic and so an over-defined set of equations should be used by increasing the amount of data. The requirement for this problem to be over-defined is,

$$n_t > \frac{n_p}{n_v} + 2.$$

7.2.3 Converting to a linear problem

As it stands, virtually all models will result in this parameter estimation problem becoming non-linear. When both a variable and a parameter occur in one term, which occurs in the majority of models, the model parameters will be multiplied by the spline coefficients producing a non-linearity. For example,

$$f(x(t)) = ax(t),$$

$$\Rightarrow f(x(t)) = a\sum_{j=0}^{n_s-1} b_j B_j(t),$$

$$\Rightarrow f(x(t)) = \sum_{j=0}^{n_s-1} ab_j B_j(t).$$

In this example there is a non-linearity between a and b_j. Unless all terms of $\mathbf{f}(\mathbf{x}(t), \gamma)$ only contain one model parameter or one model variable this problem will be non-linear.

It is desirable to have the set of equations linear so that the problem can be solved efficiently. For this to occur, the form of the model has to be restricted and an approximation made. The restriction on the model is that it must be linear in its parameters. Importantly this still allows the use of non-linear ODE models. An example of a valid non-linear model is as follows:

$$\frac{dx}{dt} = axy - bx,$$

$$\frac{dy}{dt} = cy^2.$$

Non-linearities still arise between the model parameters and the spline coefficients as described above, so an approximation is made to overcome this. An estimate vector is introduced, $\mathbf{E}_i$, which is an estimate of $\mathbf{x}(t)$ at $t = t_i$.

$$\mathbf{E}_i = \mathbf{E}(t = t_i) \approx \mathbf{x}(t_i)$$

It is assumed that the data is some noisy observation of the "real" trajectory,

$$\hat{\mathbf{x}}_i = \mathbf{x}(t_i) + \epsilon,$$

where ϵ is the amount of Gaussian distributed error. Therefore it is reasonable to assume that $\mathbf{E}_i = \hat{\mathbf{x}}_i$. As $\mathbf{E}_i$ is an estimate of the real solution (possibly a very bad estimate), it should only be used where non-linearities would occur i.e. in equation 7.2.

As $\mathbf{E}_i$ is only defined at the data time points, equation 7.2 can only be evaluated at these points. Therefore, the collocation points are set at the data time points and the collocation points are no longer taken at the B-spline nodes. So the complete algorithm that turns the parameter estimation problem into a linear algebra problem is as follows:

Algorithm 1. *Fixing $\mathbf{E}_i$, the estimate of the solution of $\mathbf{x}(t)$, to the data points ($\hat{\mathbf{x}}$), the final set of linear equations to be solved in the parameter estimation problem are:*

- *The data and spline equation*

  At each of the n_t time points that data is taken at, the following equation should be satisfied by the variables:

  $$\hat{\mathbf{x}}_i = \sum_{j=0}^{n_s-1} \mathbf{b}_j B_j(t_i). \tag{7.3}$$

- *The model and collocation equation*

  At each of the n_t time points that data is taken at, the variables should satisfy the

following equation

$$\sum_{j=0}^{n_s-1} \mathbf{b}_j B_j^{'}(t_i) = \mathbf{f}\left(\hat{\mathbf{x}}_i, \gamma\right), \tag{7.4}$$

where $\mathbf{f}()$ *is a vector of functions linear in their parameters.*

These equations can be solved using an appropriate linear algebra technique (see section 7.2.4).

Due to the change in the collocation points, the condition for an over-defined set of equations is,

$$n_t > \frac{1}{2}\left(\frac{n_p}{n_v} + n_s\right).$$

7.2.4 Solution of the linear system of equations

To solve the over-defined set of linear equations QR decomposition is used (Press *et al.*, 2002). The full set of equations can be arranged into the standard matrix notation,

$$\min |\mathbf{A} \cdot \mathbf{x} - \mathbf{b}|,$$

where $\mathbf{A}$ is an $m \times n$ matrix, $\mathbf{b}$ is a vector of length m, $\mathbf{x}$ is a vector of length n which holds the values of the model parameters and the spline coefficients. Here, m is equal to the number of equations ($2n_v n_t$) and n is equal to the number of model parameters and spline coefficients ($n_p + n_s n_v$). QR decomposition decomposes matrix $\mathbf{A}$ such that,

$$\mathbf{A} = \mathbf{Q} \cdot \mathbf{R},$$

where $\mathbf{R}$ is an upper triangular square matrix of size $n \times n$ and $\mathbf{Q}$ is orthogonal,

$$\mathbf{Q}^T \cdot \mathbf{Q} = \mathbf{I},$$

where $\mathbf{I}$ is the identity matrix. It can be shown (see appendix C.1 for proof) that the least squares solution of $\mathbf{x}$ is equivalent to solving,

$$\mathbf{R} \cdot \mathbf{x} = \mathbf{Q}^T \cdot \mathbf{b}.$$

An alternative to QR decomposition is singular value decomposition which would also highlight any redundant or nearly redundant values in $\mathbf{x}$. This is not used here due to the additional complexity of the algorithm resulting in greater processing time.

7.2.5 A demonstration of the equation matrix for a very simple example

To clarify this approach, the equation matrix will be set up for a very simple example system. Suppose that the following model is proposed,

$$\begin{aligned} \frac{dx}{dt} &= \alpha y + \beta, \\ \frac{dy}{dt} &= \mu x, \end{aligned}$$

for a system where 3 observables have been measured, $(\hat{x}_0, \hat{y}_0)$ at time t_0, $(\hat{x}_1, \hat{y}_1)$ at time t_1 and $(\hat{x}_2, \hat{y}_2)$ at time t_2. The problem is to find the parameters (α, β, μ) that will cause the model to fit the data in the best possible way. Two splines are proposed with three B-splines in each ($n_s = 3$),

$$x(t) \approx u(t) = \sum_{j=0}^{2} b_j B_j(t),$$

$$y(t) \approx v(t) = \sum_{j=0}^{2} c_j B_j(t).$$

Therefore, using Algorithm 1, the matrix equation to minimise is,

$$\begin{pmatrix}
0 & 0 & 0 & B_0(t_0) & B_1(t_0) & B_2(t_0) & 0 & 0 & 0 \\
0 & 0 & 0 & B_0(t_1) & B_1(t_1) & B_2(t_1) & 0 & 0 & 0 \\
0 & 0 & 0 & B_0(t_2) & B_1(t_2) & B_2(t_2) & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & B_0(t_0) & B_1(t_0) & B_2(t_0) \\
0 & 0 & 0 & 0 & 0 & 0 & B_0(t_1) & B_1(t_1) & B_2(t_1) \\
0 & 0 & 0 & 0 & 0 & 0 & B_0(t_2) & B_1(t_2) & B_2(t_2) \\
\hat{y}_0 & 1 & 0 & -B_0'(t_0) & -B_1'(t_0) & -B_2'(t_0) & 0 & 0 & 0 \\
\hat{y}_1 & 1 & 0 & -B_0'(t_1) & -B_1'(t_1) & -B_2'(t_1) & 0 & 0 & 0 \\
\hat{y}_2 & 1 & 0 & -B_0'(t_2) & -B_1'(t_2) & -B_2'(t_2) & 0 & 0 & 0 \\
0 & 0 & \hat{x}_0 & 0 & 0 & 0 & -B_0'(t_0) & -B_1'(t_0) & -B_2'(t_0) \\
0 & 0 & \hat{x}_1 & 0 & 0 & 0 & -B_0'(t_1) & -B_1'(t_1) & -B_2'(t_1) \\
0 & 0 & \hat{x}_2 & 0 & 0 & 0 & -B_0'(t_2) & -B_1'(t_2) & -B_2'(t_2)
\end{pmatrix}
\begin{pmatrix} \alpha \\ \beta \\ \mu \\ b_0 \\ b_1 \\ b_2 \\ c_0 \\ c_1 \\ c_2 \end{pmatrix}
-
\begin{pmatrix} \hat{x}_0 \\ \hat{x}_1 \\ \hat{x}_2 \\ \hat{y}_0 \\ \hat{y}_1 \\ \hat{y}_2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

This could be solved using QR decomposition to give estimates for the parameters but this will not be done here as the number of splines is too few to represent the solution (such a low number was used so that the matrix could easily be displayed). A more realistic example is examined below.

7.3 Tests and results

The simple four component model (model 1, equation 4.1) is used to test this procedure (see section 4.3.3). It models the behaviour of active ATM (a), p53 (z), active p53 (x)

and MDM2 (y) after a cell experiences DNA damage,

		Production	Degradation	Binding/Enzyme
$\frac{da}{dt}$	=		$-D_{ATM}a,$	
$\frac{dz}{dt}$	=	p_{p53}	$-D_{p53}z - k_1yz$	$-k_2az,$
$\frac{dx}{dt}$	=		$-D_{p53}x - k_1yx$	$+k_2az,$
$\frac{dy}{dt}$	=	$p_{MDM2} + k_3x$	$-D_{MDM2}y$	$-k_4ay.$

In total the model has nine parameters and for the purpose of this study the parameter values that the procedure will aim to recover are fixed (Table 7.1). As appropriate for this algorithm, the model is linear in its parameters but also contains non-linearities. A pseudo data set (see section 6.1.2) was created by integrating the model with the expected parameter values using an adaptive step fourth-order Runge-Kutta algorithm with high accuracy (see appendix B.5). Data sets with different numbers of time points were created from this set by sampling.

Table 7.1: The parameter values that the routine will try to recover.

D_{ATM}	D_{MDM2}	D_{p53}	p_{MDM2}	p_{p53}	k_1	k_2	k_3	k_4
0.05	0.18	0.041	0.01	0.52	1.42	0.39	2.50	0.75

7.3.1 Initial results

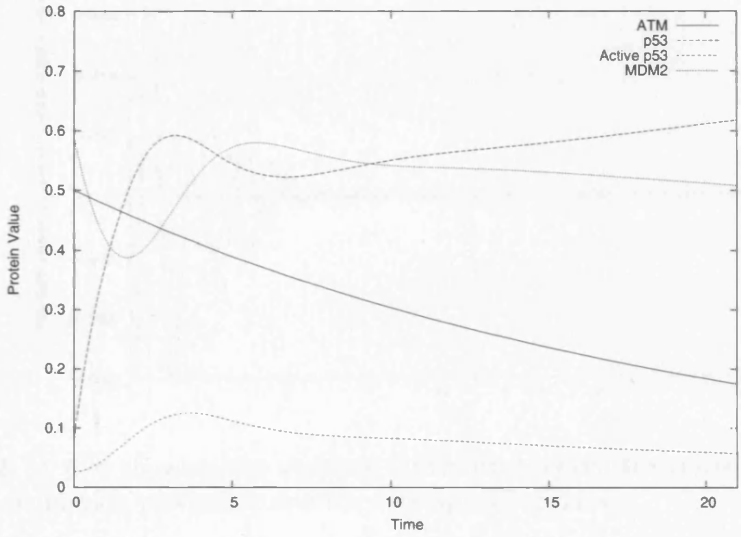
An initial test was performed on the algorithm using a data set of 1000 time points (a total of 4000 data points) and 22 B-splines. The model solution spline produced is virtually indistinguishable from the true solution (Figure 7.1). The difference between the spline and the data is very low especially in the later part of the time course (Figure 7.2). The error in the spline oscillates and this is probably because there is only a limited number of B-splines. The error rapidly decreases as time is increased. This occurs because early in the response there is rapid change making it difficult to fit the splines accurately but later on, the system is approaching equilibrium with slower changes. The error in the p53 spline is largest, probably because p53 has the most rapid change initially.

The results are very close to the desired values and are "good enough" for practical use (Table 7.2 and 7.3). It is interesting that the parameters with the two highest percentage errors are the two basal rates, D_{p53} (the basal degradation rate of p53) and p_{MDM2} (the basal production rate of MDM2). This could be because of their role in the model

Table 7.2: The parameter estimates to 6 *s.f.* from the algorithm when there are 22 B-splines and a data set containing 1000 time points per component.

D_{ATM}	D_{MDM2}	D_{p53}	p_{MDM2}	p_{p53}	k_1	k_2	k_3	k_4
0.05	0.180026	0.0406778	0.0100187	0.519501	1.41914	0.389603	2.49999	0.750031

(a)



(b)

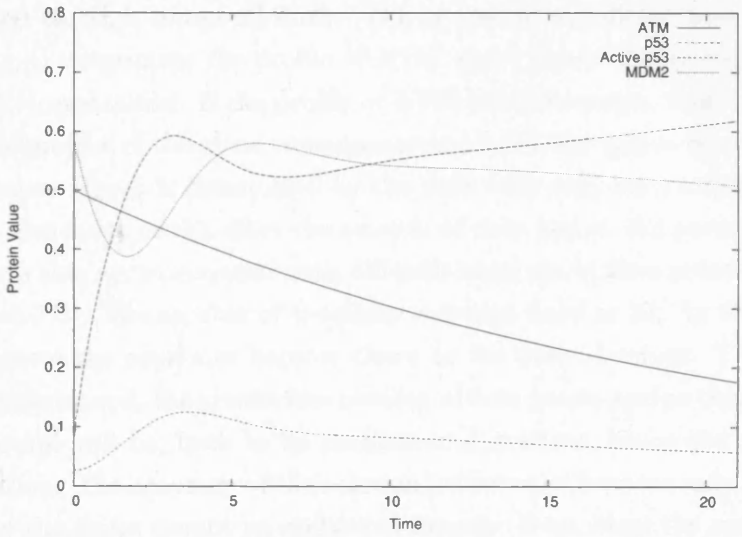


Figure 7.1: A plot showing both the (a) the spline (containing 22 B-splines) produced by the parameter estimation procedure and (b) the corresponding data which has 1000 time points.

Table 7.3: The percentage error for the parameter estimates when there are 22 B-splines and a data set with 1000 time points.

D_{ATM}	D_{MDM2}	D_{p53}	p_{MDM2}	p_{p53}	k_1	k_2	k_3	k_4
0%	0.0144%	0.786%	0.187%	0.0960%	0.0606%	0.102%	0.0004%	0.00413%

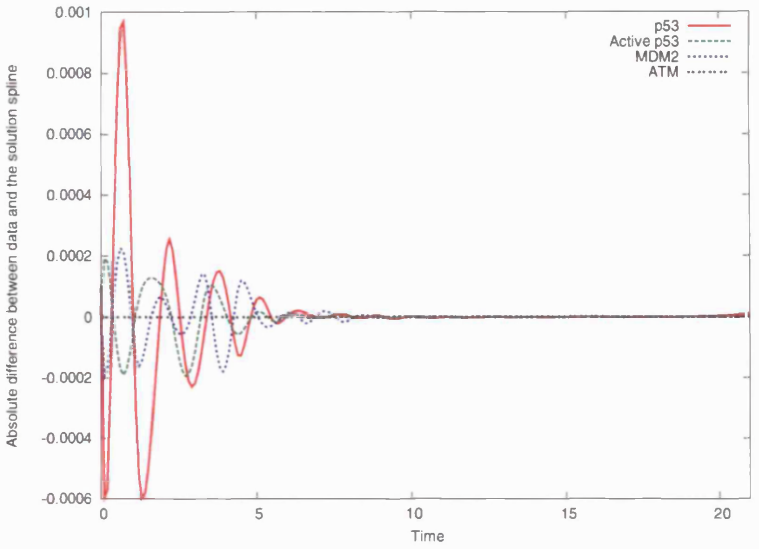


Figure 7.2: A plot showing the absolute difference between the spline produced by the parameter estimation procedure and the corresponding data.

or simply because the parameter estimates are small and so a greater amount of error is introduced through round-off error. D_{ATM} has no significant error. This could be because D_{ATM} determines the profile of ATM which drives the system so D_{ATM} needs to be heavily constrained. If the profile of ATM was inaccurate, this would have a large effect on the profiles of the other components and hence the parameter estimates. It may also be because D_{ATM} is determined by the data from only one component, ATM.

To give some idea of the effect the amount of data has on the parameter estimates, a range of data sets were processed with different numbers of time points, between 15 and 1000 (Figure 7.3). The number of B-splines was kept fixed at 22. As the number of time points increases the estimates become closer to the desired values. This is because the more time points used, the greater the number of data points and so the more constrained the time course will be, both in its position and gradient, hence the closer it is to the "true" solution. The accuracy of the solution is restricted because only 22 B-splines were used, so the dynamics cannot be replicated exactly. Even when the amount of data was small the parameter estimates were still reasonably good, and definitely good enough to be used as the seed for another parameter estimation routine.

For all parameters, the size of error evolves in a similar way; at a low number of time points, a small increase in their number causes a very rapid improvement in the error, but as the number of time points increases the amount of improvement decreases. When working with small amounts of data, which is likely in this project, this suggests that even one additional experiment can cause a large improvement in the quality of the parameter estimates. For most parameters, the estimates seem to converge to an insignificant error at large amounts of data, but there are a few (D_{p53}, p_{p53}, k_1 and k_2) where the convergence is very slow or converging to a non-zero value.

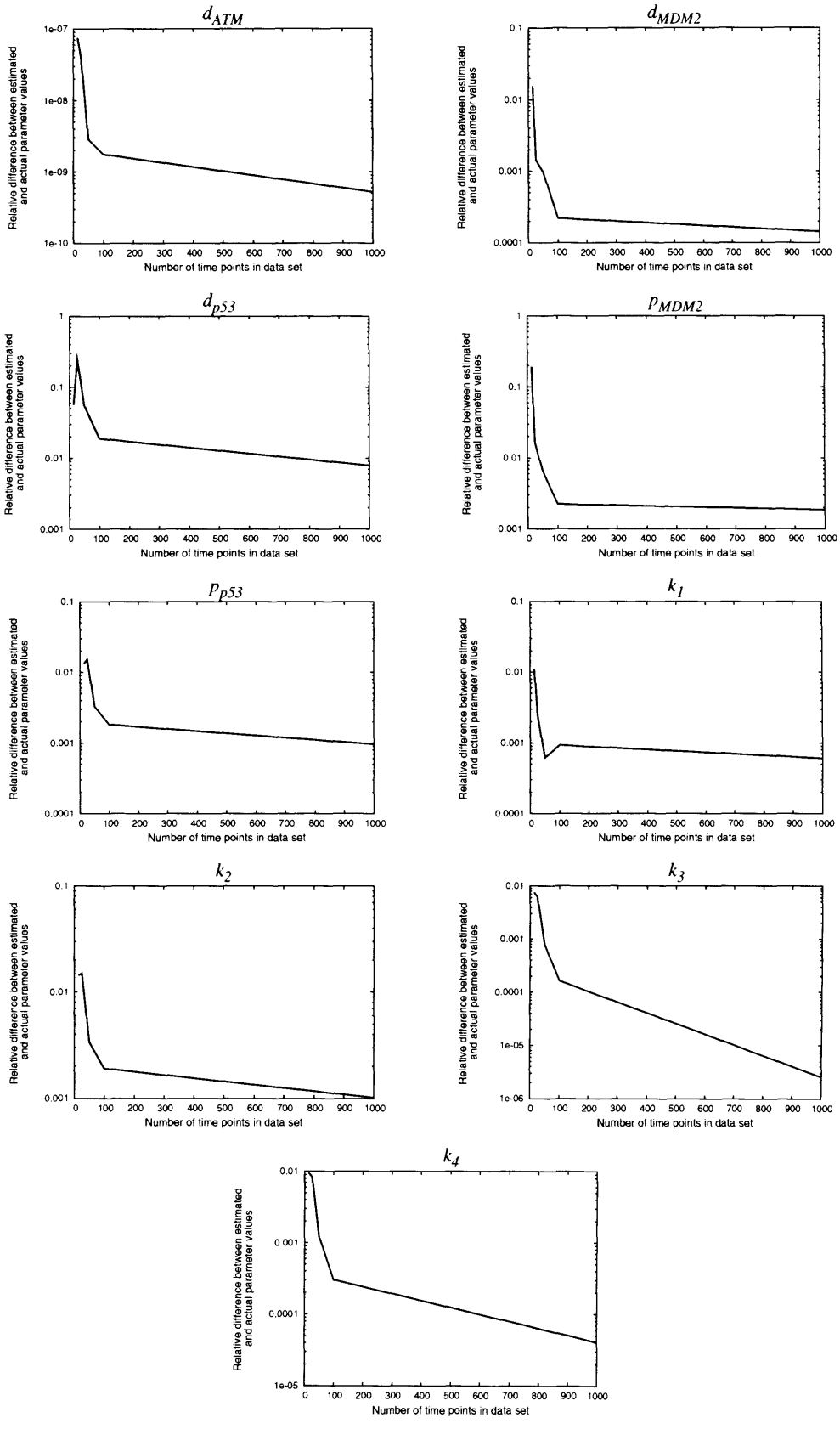


Figure 7.3: A plot showing how the parameter estimation improves with increasing number of time points in the data set. The number of B-splines is set to 22.

To examine the effect the number of B-splines has on the parameter estimates, a data set of 1000 time points was processed with a range of different numbers of B-splines from 7 to 100 (Figure 7.4). As the number of B-splines is increased the spline's ability to represent the profiles of the components is also increased and so the error in the parameter estimate decreases. Again the amount of improvement decreases with increasing number of B-splines. For the majority of parameters 22 B-splines are enough to produce reasonably good estimates (less than 0.001 relative error). These results suggest that the number of B-splines should be set as large as possible but increasing the number of B-splines also increases the minimum data requirement and slows the computation of the solution.

For all of the parameter estimates the error becomes insignificant with enough B-splines. Different parameter estimates converge to a good estimate (10^{-6}) at different rates. The fastest seems to be D_{ATM} followed by k_3 and k_4. D_{ATM} controls the component that drives the system and so its accuracy is important. k_3 and k_4 are the only non-basal parameters that affect MDM2 concentration and so the rapid convergence of k_3 and k_4 might indicate that MDM2 is the secondary driver of the system. Parameters p_{p53}, D_{p53}, k_1 and k_2 seem particularly slow to reach a good estimate. This set overlaps with the slowly converging set found when examining how the amount of data affects the estimates. Maybe this is indicative that these parameters have less affect on the behaviour of the system.

This parameter estimation method works extremely well, even at reasonably low numbers of time points and low numbers of collocation parameters. The algorithm runs quickly and is very reliable when compared to the methods examined in chapter 6.

7.3.2 Adding noise to the data

To be of practical use the algorithm needs to be able to cope well with realistic data that has error. To simulate error in the data, each data point was sampled from a Gaussian with the mean set at the true value and the variance set at σ^2, where σ is a constant value for all data points. The parameter estimation algorithm was then run on this new data set. This was repeated 1000 times at a range of different σs from 0 to 0.06. A σ of 0.06 is approximately 75% error for active p53 and 12% error for the other components. The number of time points in the data was set to 106 and the number of B-splines set to 22.

The parameter estimates diverge rapidly from the true values as the error in the data is increased (Figure 7.5). If the parameter estimation method was working well it is expected that the average would be near to the true value (assuming that the parameter estimates are distributed as a Gaussian). After 0.005 error is added to the data the majority of the average parameter values are not within one standard deviation of the true value, indicating that the parameter estimates are not in agreement with the true

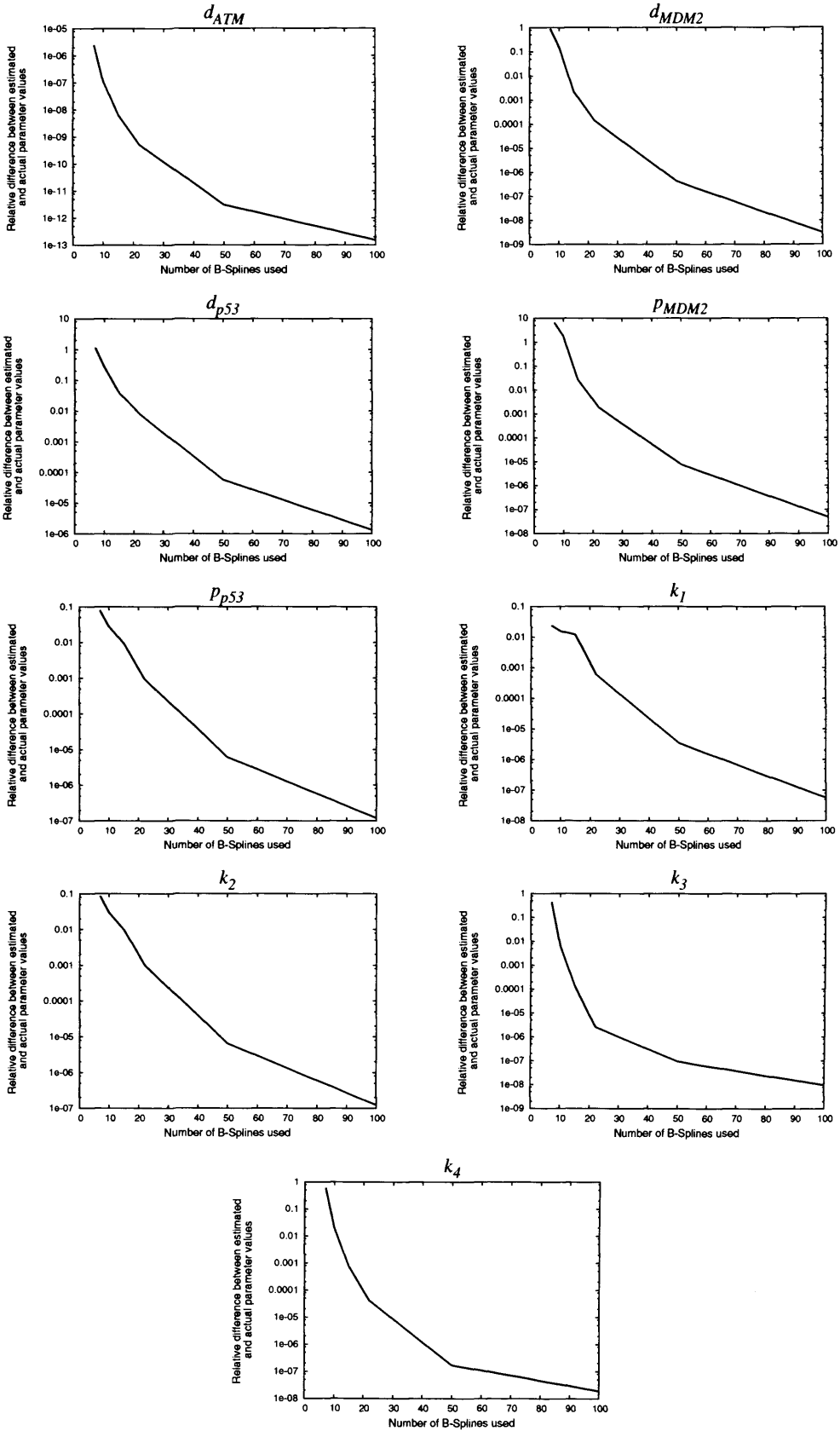


Figure 7.4: A plot showing how the parameter estimation improves with increasing number of B-splines. The number of time points is set to 1000.

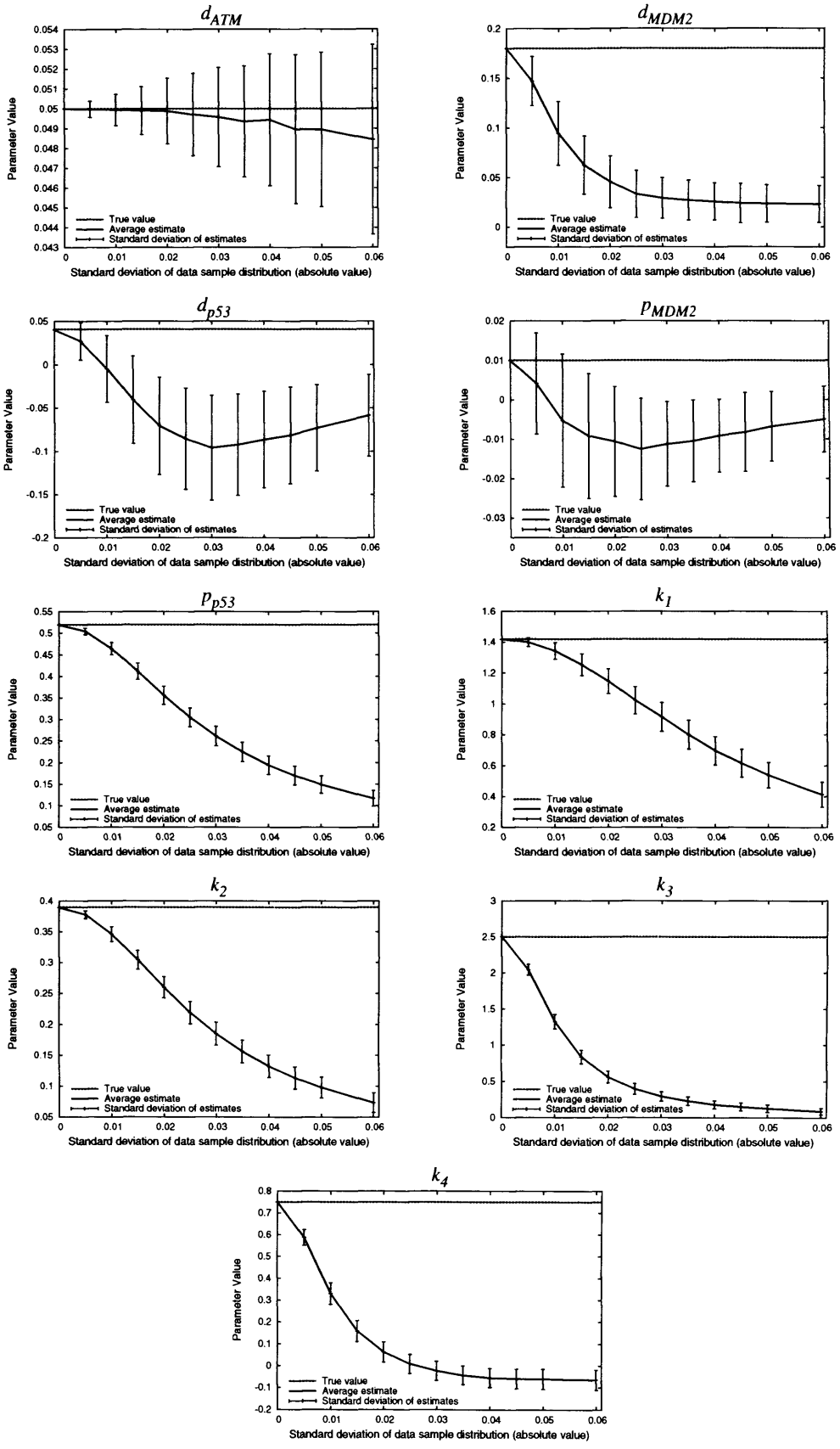


Figure 7.5: A plot showing how the average parameter estimate changes with error in the data (based on 1000 runs). The number of B-splines is set to 22 and the number of time points is 106. The error bars show the first standard deviation of the estimates.

values[1]. Exceptions to this are D_{ATM}, p_{MDM2} and D_{p53}. The average parameter value of D_{ATM} is close to the expected value with a relatively low "error". D_{p53} and p_{MDM2} though, have a large "error" and the averages rapidly diverge from the true values. This suggests that D_{p53} and p_{MDM2} are difficult to discern and do not have a large effect on the dynamics of the system. Algorithm 1 does not appear to work well with data that has even a small amount of error.

7.4 An improvement on the algorithm

As the amount of error in the data is increased the parameter estimates predicted by Algorithm 1 rapidly become inaccurate. The reason for this is that when there is error in the data, using it as an estimate of the solution, $\mathbf{E}_i$, is very inaccurate (see section 7.2.3). An improvement to Algorithm 1 would be to repeatedly solve the problem by QR decomposition updating $\mathbf{E}_i$ based on the previous result.

Algorithm 2. *An iterative method of parameter estimation using collocation and linear algebra would be to repeatedly run Algorithm 1 with a changing estimate of the solution,* $\mathbf{E}_i$. $\mathbf{E}_i$ *is set equal to the approximated solution* $\mathbf{u}(t)$ *from the previous iteration i.e.*

$$\begin{aligned}
\mathbf{E}_{i0} &= \hat{\mathbf{x}}_i, \\
\mathbf{E}_{i1} &= \sum_{j=0}^{n_s-1} \mathbf{b}_{j0} B_j(t_i), \\
&\vdots \\
\mathbf{E}_{in} &= \sum_{j=0}^{n_s-1} \mathbf{b}_{j(n-1)} B_j(t_i) = \mathbf{E}_{i(n-1)},
\end{aligned}$$

where $\mathbf{E}_{iq}$ *is* $\mathbf{E}_i$ *at the qth iteration and* $\mathbf{b}_{jq}$ *are the spline coefficients at the qth iteration. For the first iteration* $\mathbf{E}_i$ *would be set equal to the data points as before. This algorithm has converged if the estimates from one iteration to the next are the same within some tolerance. Once it has converged equation 7.2 has in effect been solved because the estimate equals the solution i.e.* $\mathbf{E}_i = \mathbf{u}(t_i) \approx \mathbf{x}(t_i)$.

7.4.1 Comparison with Algorithm 1

As in section 7.3.2 the data was resampled with added error and Algorithm 2 was applied. This was repeated 1000 times at a number of different amounts of error. The average parameter estimate of Algorithm 2 diverges at a slower rate than Algorithm 1, indicating that the iterative method is a considerable improvement (Figure 7.6). Previously, some mechanisms were removed at high levels of error, for example k_3 and k_4 had estimates

[1] 68% of the parameter values lie within one standard deviation (if the assumption of a Gaussian distribution holds) and so can be used as a coarse measurement of error for the average parameter value.

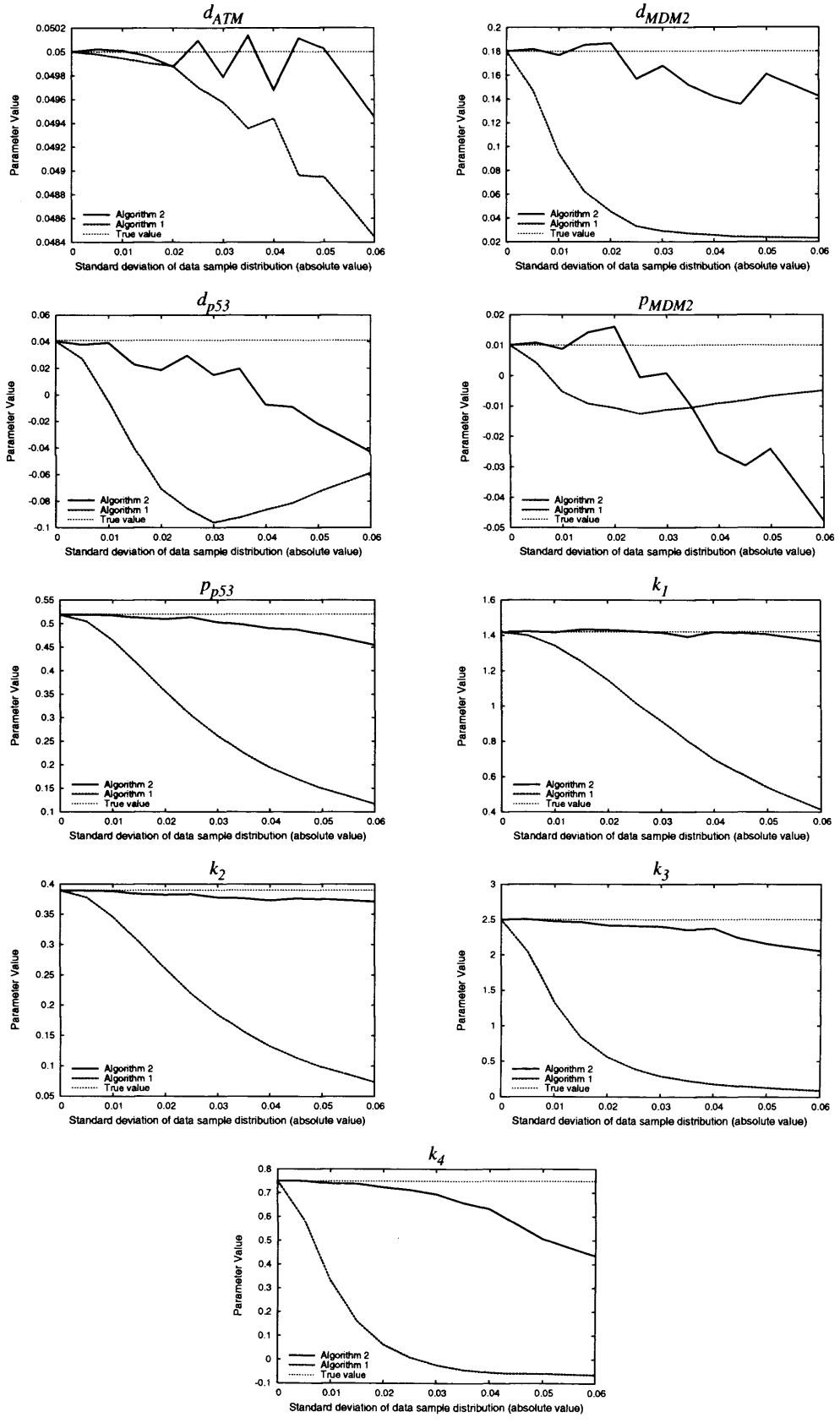


Figure 7.6: A plot comparing the average parameter estimate (based on 1000 runs) of Algorithm 1 and Algorithm 2 (the iteration version). 22 B-splines and the 106 time point data set was used.

close to zero. In Algorithm 2 these estimates remain close to their true value. Algorithm 2 still does diverge and it is unclear why this occurs. Even though the error distribution is Gaussian it is not necessarily true that the parameter estimates will be distributed as a Gaussian especially at large values of error. The amount of divergence though makes it seem likely that the algorithm is not performing optimally at large values of error.

The true parameter values are well within the first standard deviation of the distribution of parameter values for all parameters (Figure 7.7). This means that the average parameter estimate and the true value are in agreement if the coarse error estimate is used. Additionally it looks like the standard deviation is approximately proportional to the error introduced into the data. This is encouraging as it appears that the error is being propagated through the algorithm. The standard deviations are normally larger than with Algorithm 1; this suggests that there is a wide range of parameter values one can achieve using this method, which implies that in individual cases the parameter estimates can be distant from the true values. The difficulty is knowing what "correct" means in this context; if in each individual case it finds the best parameter values based on the available data, then that is all that is required.

7.4.2 Comparison with a non-linear error function

If this technique is working well similar solutions should be obtained when using a non-linear error function without the use of splines. Using error adapted data, Algorithm 2 was run and the results used as the initial point of the Nelder-Mead method (Nelder and Mead, 1965). An adaptive fourth order Runge Kutta method run with a high accuracy was used to integrate the model equations. The initial conditions were set at the true initial values. This was repeated 1000 times for a variety of data sets with different amounts of error.

The average difference between the parameter estimates of the two methods increases with increasing error in the data (Figure 7.8). At small amounts of error the difference is small, indicating that in this situation the algorithm is working well. The amount of divergence varies significantly between parameters. For p_{MDM2}, the largest difference is nearly 600% while for D_{ATM} the largest difference is about 10%. Two out of the three parameters with highest divergence are the basal rates which have been previously suggested to be problematic to estimate. Again, D_{ATM} is suggested as an important factor as it has the lowest divergence.

To quantify how well each method produces the true parameter values a *parameter fit measure* is introduced,

$$A = \sum_{i}^{n_p} \left(\gamma_i - \gamma_i^T\right)^2, \tag{7.5}$$

where γ_i is the estimate of the parameter i produced by either parameter estimation routine and γ_i^T is the true parameter value. On average the simplex method produces

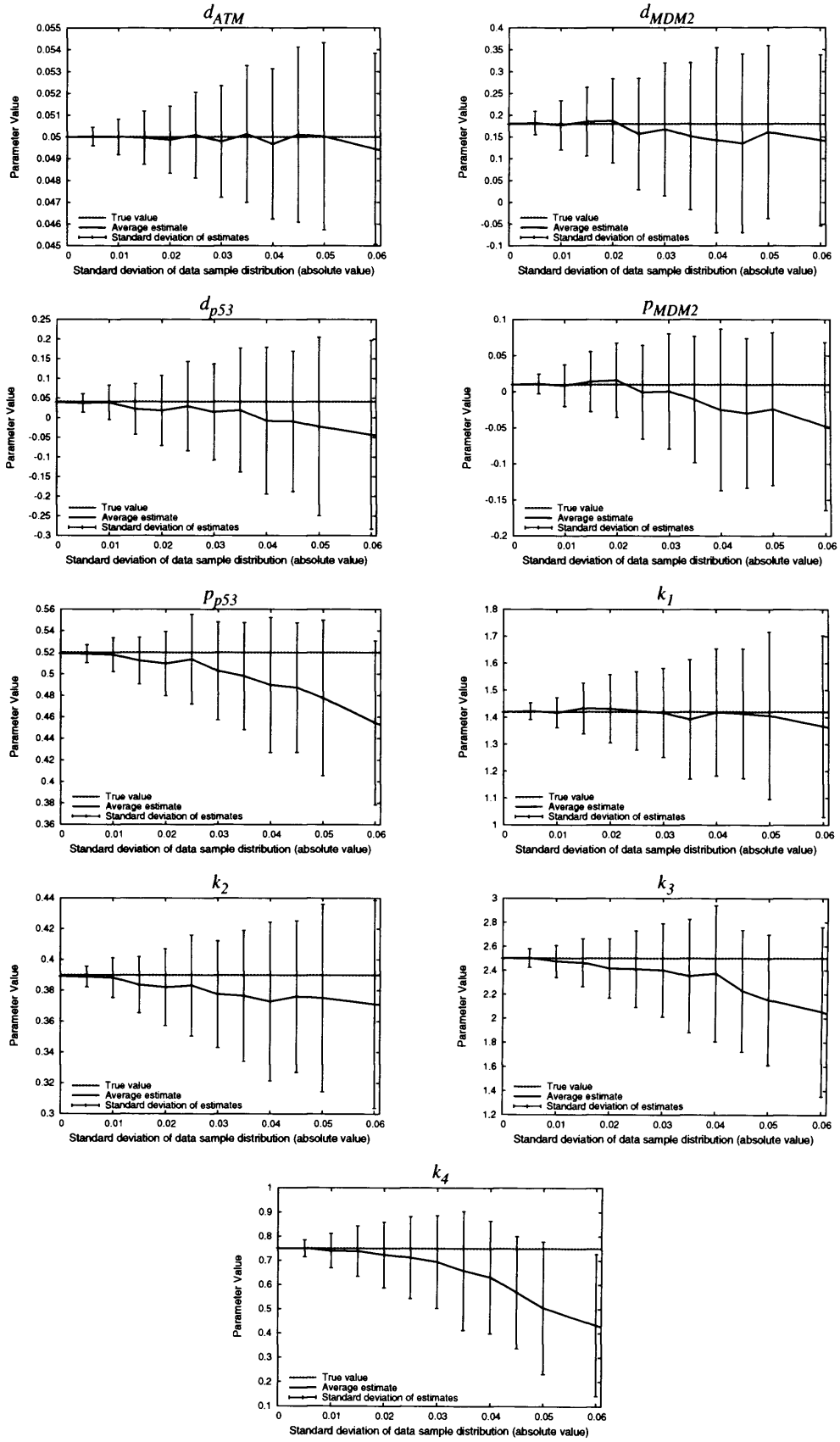


Figure 7.7: A plot showing how the average parameter estimate changes with error in the data (based on 1000 runs) when using Algorithm 2. 22 B-splines and the 106 time point data set was used.

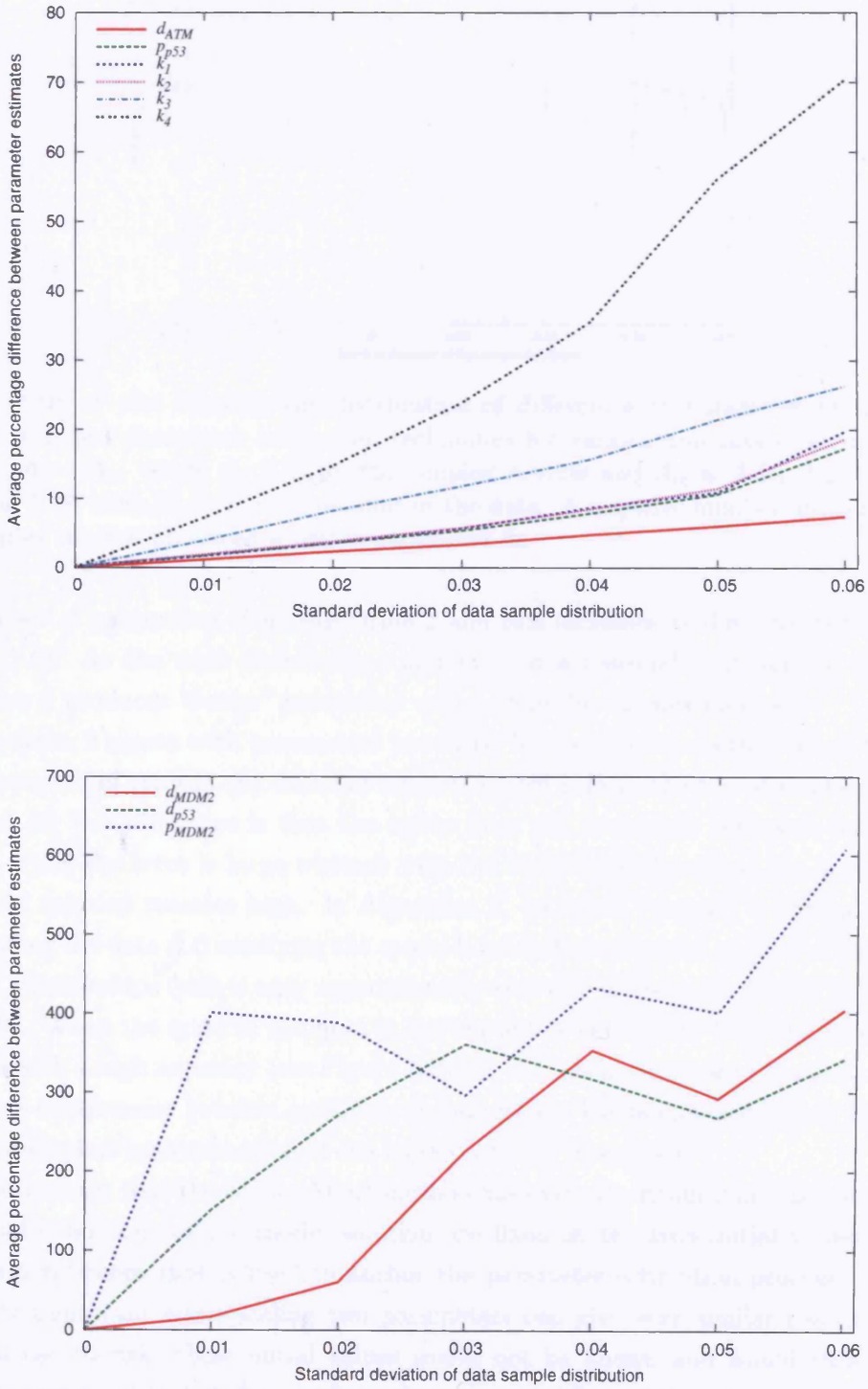


Figure 7.8: A plot showing how the average percentage difference between the estimated parameters of Algorithm 2 and the simplex routine varies with increased error in the data set. Average based on 1000 trials.

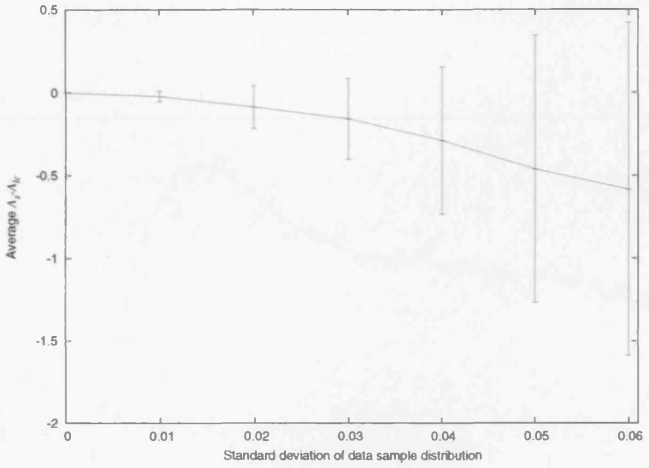


Figure 7.9: A plot showing the distribution of differences in parameter fit measure between the two parameter estimation techniques for various amounts of error in the data i.e. $A_s - A_{lc}$, where A_s is A for the simplex routine and A_{lc} is A for Algorithm 2. Based on 1000 trials at each error amount in the data. A negative number indicates that the simplex routine produced a better parameter fit.

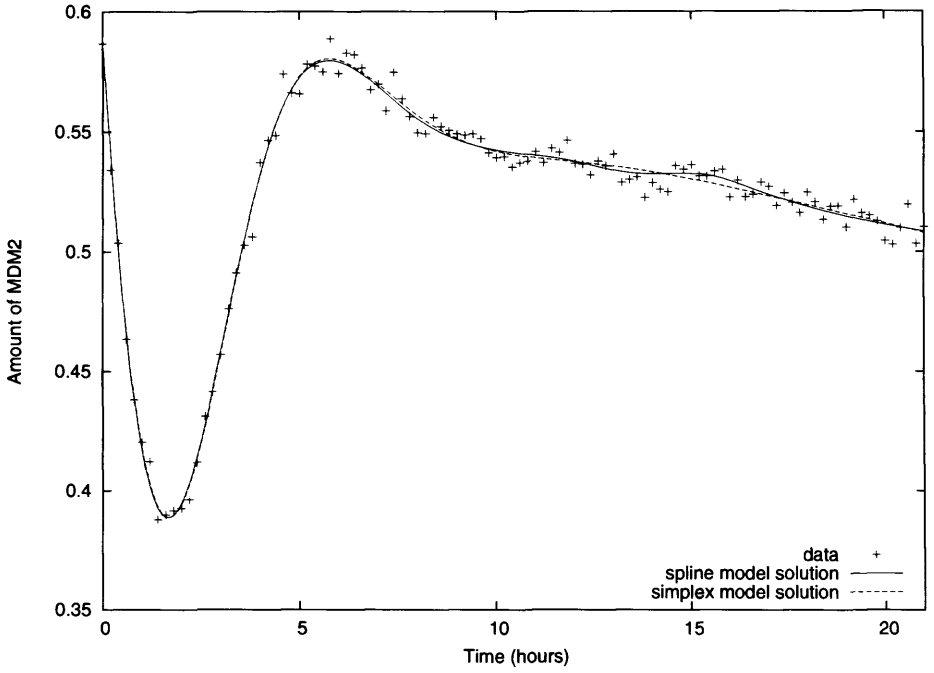
a better set of parameters than Algorithm 2 and this increases as the error is increased (Figure 7.9). As the wide distributions indicate, for a reasonable amount of the runs Algorithm 2 produces "better" parameter values than the simplex method.

Algorithm 2 agrees with parameters produced by the simplex method when there is a small amount of error in the data but when the error is large the two solutions diverge. A reason for this difference is that the spline does not accurately represent the model solution when the error is large whereas with the Nelder-Mead method the accuracy of the model solution remains high. In Algorithm 2, an equal "weight" is applied on the spline fitting the data and satisfying the model because there is equal number of equations on both relationships (this is only approximately true due to the scale of each individual equation). When the error in the data is low the spline can satisfy both the model and the data with a high accuracy (see Figure 7.10(a)) but when the error in the data is high it makes a compromise between satisfying the model and the points (see Figure 7.10(b)), making the spline approximation of the model solution inaccurate.

An advantage that the Nelder-Mead method has over Algorithm 2 in this test is that the initial conditions of the model solution are fixed at the true initial values. This provides a reference that is used to anchor the parameter estimation process. This is especially important when scaling two parameters can give very similar results[2]. For practical use though, these initial values would not be known and would either have to be guessed at of included as additional parameter. One of the advantages of Algorithm 2 is that information of the initial conditions is not required. Another advantage of Algorithm 2 is that it is less computationally intensive.

[2]This is a possible reason why there is such a large difference between the estimates for p_{MDM2} and D_{MDM2}

(a)



(b)

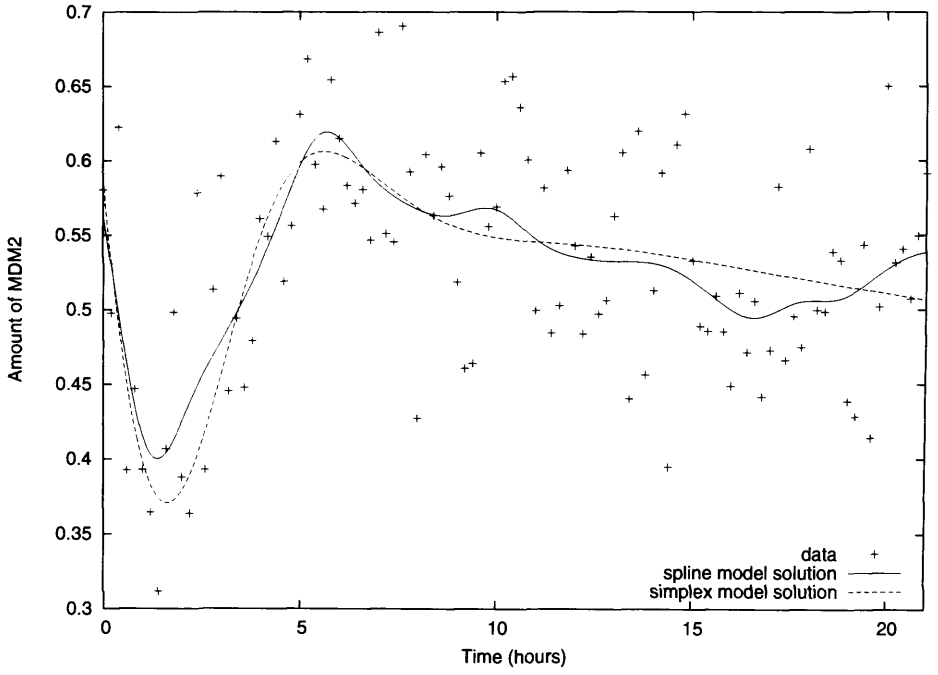


Figure 7.10: The model solution for MDM2 when using the estimated parameters from one run of Algorithm 2 and the simplex method for (a) 0.01 error in the data and (b) 0.06 error in the data.

7.5 Quantifying the errors in the parameters

For a parameter estimation technique to be useful the algorithm should not only find the parameters that give the best fit but also give error estimates on the parameters (Press *et al.*, 2002). Here estimates of the error in the parameters are examined. Consider the parameter estimation problem in matrix notation,

$$\min |\mathbf{A} \cdot \mathbf{x} - \mathbf{b}| .$$

Each value of b_i has an error associated with it, σ_i. In equation 7.1, σ_i is the error associated with each experimental measurement and for equation 7.2, σ_i is some measure of how accurate the estimated solution, $\mathbf{u}(t)$, is to the real solution, $\mathbf{x}(t)$. For equation 7.2, σ_i might be particularly difficult to quantify. If the equations are assumed to be independent, then each equation contributes a part of its own uncertainty to the parameters. Therefore the variance in the value of a parameter, γ, can be quantified as follows (Press *et al.*, 2002),

$$\sigma_\gamma^2 = \sum_{i=1}^{2n_d n_v} \sigma_i^2 \left(\frac{\partial \gamma}{\partial b_i} \right)^2 .$$

Applying this to the QR decomposition algorithm it can be shown that,

$$\sigma^2(\gamma_j) = \sum_{i=1}^{2n_d n_v} \sigma_i^2 \left(\sum_{k=1}^{n_p+n_s} R_{jk}^{-1} Q_{ki}^T \right)^2 ,$$

where γ_j is the jth parameter and G_{jk} is the element of matrix $\mathbf{G}$ in the jth row and kth column (see appendix C.2 for proof). Other possible ways of getting a measure of error on the parameter estimate is to use MCMC to get a probability distribution of the parameter value, use the Hessian matrix, or a bootstrap method (Press *et al.*, 2002).

As an example of this error measurement the σ values were all set to 0.03 and error from a Gaussian distribution with a standard deviation of 0.03 was added to the data set. It is assumed that there is no error associated with the second set of equations (equation 7.2). From the results of three runs, it is clear that the errors calculated are not large enough to place the expected values and the estimated parameter values in agreement (Table 7.4). This could be because no error was associated with half of the equations. Alternatively it could be to do with the conception of what the best parameter values are with these data sets, for example there is a certain probability that the data set will have flat profiles, in this case there is no chance that the parameter values will be correct even though the data has the same error. On a couple of equations the parameter values given are negative; there is no guarantee that this parameter estimation method will produce positive values. Again the small errors for D_{ATM} indicate the importance of getting this parameter correct.

Table 7.4: The parameter estimates and errors for 3 runs on a 106 time point data set with 0.03 error and 22 B-splines.

	D_{ATM}	D_{MDM2}	D_{p53}	p_{MDM2}	p_{p53}	k_1	k_2	k_3	k_4
Expected values	0.05	0.18	0.041	0.01	0.52	1.42	0.39	2.50	0.75
Run 1	0.05072	0.2971	0.1562	0.08871	0.5047	1.145	0.3834	2.343	0.7584
Run 1 error	9.3×10^{-5}	0.0061	0.014	0.0017	0.0021	0.021	0.0016	0.089	0.020
Run 2	0.05235	0.1330	0.083400	−0.01433	0.5862	1.552	0.4248	2.061	0.4788
Run 2 error	9.2×10^{-5}	0.0057	0.0094	0.0016	0.0021	0.020	0.0014	0.062	0.011
Run 3	0.04937	0.2595	−0.005642	0.08892	0.4629	1.362	0.3395	2.138	0.7780
Run 3 error	9.2×10^{-5}	0.0071	0.012	0.0020	0.0021	0.023	0.0014	0.064	0.017

7.6 A further refinement to the algorithm

7.6.1 Introduction - the problem

In Algorithm 2 there are two types of equations: one set that examines how close the spline is to the data (data-spline equations, equation 7.1) and another set that examines how close the spline is to the model solution (model-spline equations, equation 7.2). The algorithm spline acts as an intermediary between the data and the model, attempting to be as close to each as possible. Let r_{data} be the sum of the squared residuals of the data-spline equations,

$$r_{\text{data}} = \sum_{i \in \xi} r_i^2,$$

where ξ is the set of data-spline equations and r_i is the residual from a single equation. Similarly, r_{model} is the sum of the squared residuals of the model-spline equations. The solution parameters are those that minimise $r_{\text{data}} + r_{\text{model}}$, which occurs when $r_{\text{data}} = r_{\text{model}}$[3]. In most biological measurements the error in the data is likely to be reasonably high, this means that the r_{data}, and hence r_{model} will probably be reasonably high too. In effect the method manifests a large difference between the model and data by producing a spline that is a compromise between the data and model, not fitting the data nor the model that well. Compared to the true model solution the spline produced when there is a large amounts of error "wiggles around" and has a large r_{data} and r_{model} (Figure 7.11).

This type of behaviour is problematic because if the spline does not fit the model well the parameter estimates are unlikely to be accurate. It is much more important that the spline is a good approximation to the model solution rather than the data because it is unlikely that the spline should go exactly through the data points because of the experimental error whereas the spline should perfectly represent the model.

7.6.2 Increasing the number of collocation points

Changing the method so that the spline is a better approximation to the model solution will produce better parameter estimates. One way to increase the weight that is placed

[3]In truth, this would only occur if the equations were scaled to have the sum of the prefixes and constants as the same value, but is used here for illustrative reasons.

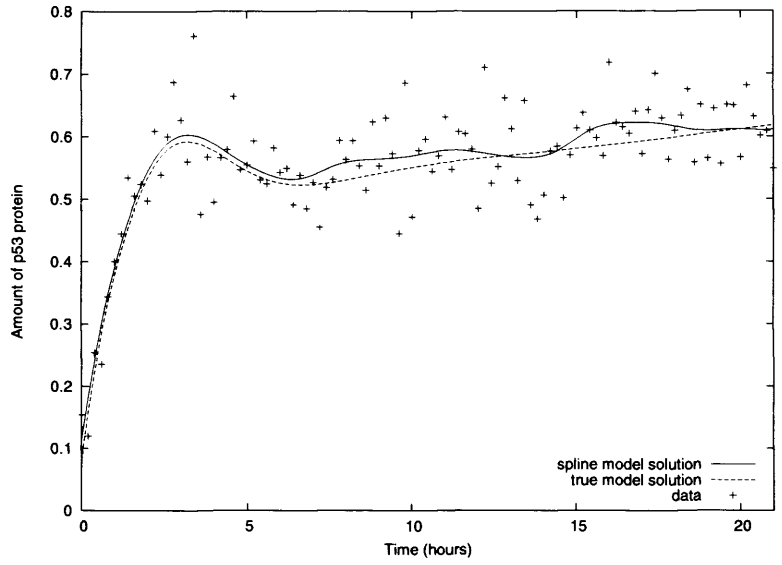


Figure 7.11: An example of the spline produced when there is 0.6 error in the data. This is using the simple model and the trace of p53 is shown. There are 106 data points and 22 B-splines in the spline. Here $r_{\text{data}} = 1.29$ and $r_{\text{model}} = 0.0493$. When there is no error in the data $r_{\text{data}} = 7.84 \times 10^{-6}$ and $r_{\text{model}} = 2.05 \times 10^{-4}$.

on the spline accurately representing the model is to increase the number of equations in the model-spline set (equation 7.2) by increasing the number of collocation points.

In Algorithm 2 the estimate vector $\mathbf{E}_i$ restricts the collocation points to the data points. To get around this restriction the estimate vector will be changed to an estimate spline, which is an approximation to the model solution,

$$\mathbf{E}(t) = \sum_{j=0}^{n_s-1} \mathbf{b}_j^* B_j(t),$$

where $\mathbf{b}_j^*$ is a vector of n_v constants associated with the jth B-spline. For convenience, $\mathbf{E}(t)$ has the same number of B-splines as used to estimate the model solution, $\mathbf{u}(t)$; these two model estimate splines can be regarded as the same. As with the estimate vector, the estimate spline is initially constructed from the data by fitting $\mathbf{E}(t)$ to the data. For this to be exactly solvable the number of B-splines, n_s, must be less than or equal to n_t. If $n_t < n_s \leq n_t + 2$, then the spline must also satisfy the condition that the two extreme nodes will have their second differentials equal to zero. This is the natural spline assumption and creates a spline with minimal curvature (Golub and Ortega, 1992).

Let there be n_c collocation points spaced equally between the first and last time point. The position of the pth collocation point is at $t = t_p^c = t_0 + p \times \frac{t_{n_t} - t_0}{n_c - 1}$. The altered algorithm is as follows:

Algorithm 3. *An iterative method of parameter estimation using a spline rather than a vector as the estimated solution*

1. *Fit the estimate spline,* $\mathbf{E}_0(t)$*, to the data,* $\hat{\mathbf{x}}_\mathbf{i}$.

2. *Solve the following equations:*

   - *The data and spline equation*

     At each of the n_t *time points that data is taken at, the following equation should be satisfied by the variables:*

     $$\hat{\mathbf{x}}_i = \sum_{j=0}^{n_s-1} \mathbf{b}_j B_j(t_i). \tag{7.6}$$

   - *The model and the collocation equation*

     At each of the n_c *collocation points,* $t_0^c \ldots t_p^c \ldots t_{n_c-1}^c$*, the variables should satisfy the following equation*

     $$\sum_{j=0}^{n_s-1} \mathbf{b}_j B_j'(t_p^c) = \mathbf{f}\left(\mathbf{E_q}(t_p^c), \gamma\right), \tag{7.7}$$

     where $\mathbf{f}()$ *is a vector of functions linear in their parameters and* $\mathbf{E_q}$ *is the estimated spline for the* q*th iteration.*

3. *The estimated spline,* $\mathbf{E}(t)$*, is updated to equal the solution spline and the equations are solved again i.e.*

$$\begin{aligned}
\mathbf{E}_1(t) &= \sum_{j=0}^{n_s-1} \mathbf{b}_{j0} B_j(t), \\
\mathbf{E}_2(t) &= \sum_{j=0}^{n_s-1} \mathbf{b}_{j1} B_j(t), \\
&\vdots \\
\mathbf{E}_n(t) &= \sum_{j=0}^{n_s-1} \mathbf{b}_{j(n-1)} B_j(t_i) = \mathbf{E}_{n-1}(t),
\end{aligned}$$

   where $\mathbf{E}_q(t)$ *is* $\mathbf{E}(t)$ *at the* q*th iteration and* $\mathbf{b}_{jq}$ *are the solution spline coefficients found at the* q*th iteration. This is repeated until* $\mathbf{E}_n(t) = \mathbf{E}_{n-1}(t)$ *within some tolerance. Once converged* $\mathbf{E}(t) = \mathbf{u}(t) \approx \mathbf{x}(t)$.

For the equations to be solvable the following condition must hold,

$$n_t + n_c \geq \frac{n_p}{n_v} + n_s.$$

Using this adaptation reduces the restriction on the number of collocation points and so has the ability of "re-weighing" the sets of equations in favour of the spline being close to

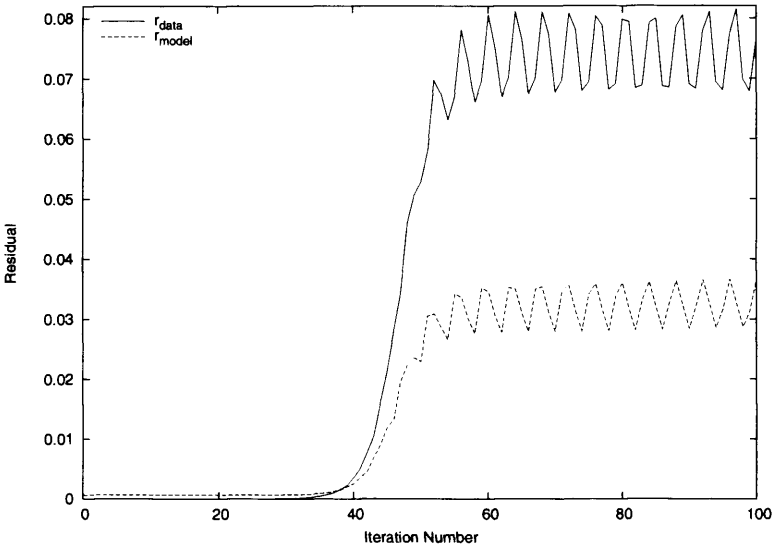


Figure 7.12: A plot showing the divergent behaviour that emerges if the number of collocation points is too high (here set at 500). This is based on the example model (see section 4.3.3), 106 time points in the data set with no error and 22 B-splines. The residual appears to increase exponentially before settling to a stable oscillation.

the model. A beneficial side effect of this change is that the data points no longer need to be regularly spaced, making the algorithm more flexible, but this might have unforeseen effects on the effectiveness of the algorithm.

7.6.3 Initial Results

As the principal aim of this modification is to produce splines that are more accurate estimations of the model solution, a model score, Y, is proposed, where,

$$Y = \sum_{i=0}^{n_v-1} \int_{t_0}^{t_{n_t-1}} \left(\sum_{j=0}^{n_s-1} b_{ji} B_j'(t) - f_i \left(\sum_{j=0}^{n_s-1} b_{ji} B_j(t), \gamma \right) \right)^2 dt, \tag{7.8}$$

where b_{ji} is the ith component of $\mathbf{b}_j$ and $f_i()$ is the ith component of $\mathbf{f}()$. This is evaluated using Simpson's rule with 10000 strips (Kreyszig, 1993). Y gives a score of how well the model is satisfied by the spline across its whole range, the lower the score the better.

Using Algorithm 3 it soon became apparent that it did not behave correctly if there were too many collocation points. A distinct pattern emerged: the residual exponentially increases away from the initial value until it reached an equilibrium (Figure 7.12), and the resulting estimated parameters are inaccurate (see Table 7.5). In this situation, the Y-score and residuals are poor; when the algorithm is applied to the example model with 106 time points in the data set, 22 B-splines and 500 collocation points the Y-score was 0.0164 and the sum of the residuals 0.0991 (the corresponding scores with Algorithm 2 when $n_c = 106$ are 3.22×10^{-5} and 2.12×10^{-4}). When this situation occurs the run will

Table 7.5: The parameter estimates for Algorithm 2 (106 collocation points, convergent behaviour) and Algorithm 3 with 500 collocation points (divergent behaviour). There are 22 B-splines and a data set containing 106 time points per component with no error.

n_c	D_{ATM}	D_{MDM2}	D_{p53}	p_{MDM2}	p_{p53}	k_1	k_2	k_3	k_4
106	0.05	0.18	0.041	0.01	0.519	1.42	0.389	2.50	0.75
500	0.05	0.275	−0.135	0.0540	0.363	1.29	0.268	2.70	0.807

be described as being divergent or if not the run will be described as convergent[4].

Various changes were made to the algorithm to ensure that this behaviour was not due to an implementation detail. The linear solver was changed to a singular value decomposition (SVD) solver (Press *et al.*, 2002) and the divergent behaviour was still present. SVD also provides information about the condition of the matrix. In the divergent example case of 500 collocation points the matrix is well conditioned with a condition number[5] of about 225. Algorithm 3 was also applied to a damped oscillator model and adapted in a number of different ways such as solving the problem in two steps, setting the number of B-splines equal to the number of collocation points and using the spline produced by Algorithm 2 as the initial spline for Algorithm 3 (see Appendix C.3). In all cases the divergent behaviour persisted. A method was devised to quickly detect whether a run was divergent so that computational resources could be saved (see Appendix C.4).

For a divergent run the solution spline is moving away from the data points and becoming a worst estimate of the model solution. After each iteration the estimated spline for the next iteration is set equal to the solution spline found from the current iteration. This update to the estimate spline is the link between iterations and so this process might be causing some sort of positive feedback allowing small "errors" to propagate and amplify. The algorithm diverges when there is a greater number of collocation points and hence more weight placed on the spline satisfying the model than the spline satisfying the data. This leads to the hypothesis that there can be too much weight placed on the spline satisfying the model. One possibility is that the data acts as an anchor for the spline, if too much weight is placed on the spline satisfying the model then the data does not have enough hold on the spline's position allowing the spline to move away.

For the example model and data set when there is no error the change between convergent and divergent behaviour occurs when $n_c \approx 295$. As the number of collocation points increases it takes a greater number of iterations before convergence (Figure 7.13(a)), approaching infinity at the threshold between convergent and divergent behaviour. Despite the increased processing time, there are benefits in having a higher number of collocation points, the Y-score and the average residual per equation for both sets of equations decrease with increasing collocation number (Figures 7.13(b)–7.13(d)). Most significant

[4]These terms are based on the initial behaviour, in both cases they are finally convergent with respect to the residual, but when it is divergent the residual normally stabilises to an oscillation at a high level

[5]If the reciprocal of the condition number reaches floating-point precision, then numerically solving the matrix can be problematic, but the matrices are well away from this condition (Press *et al.*, 2002).

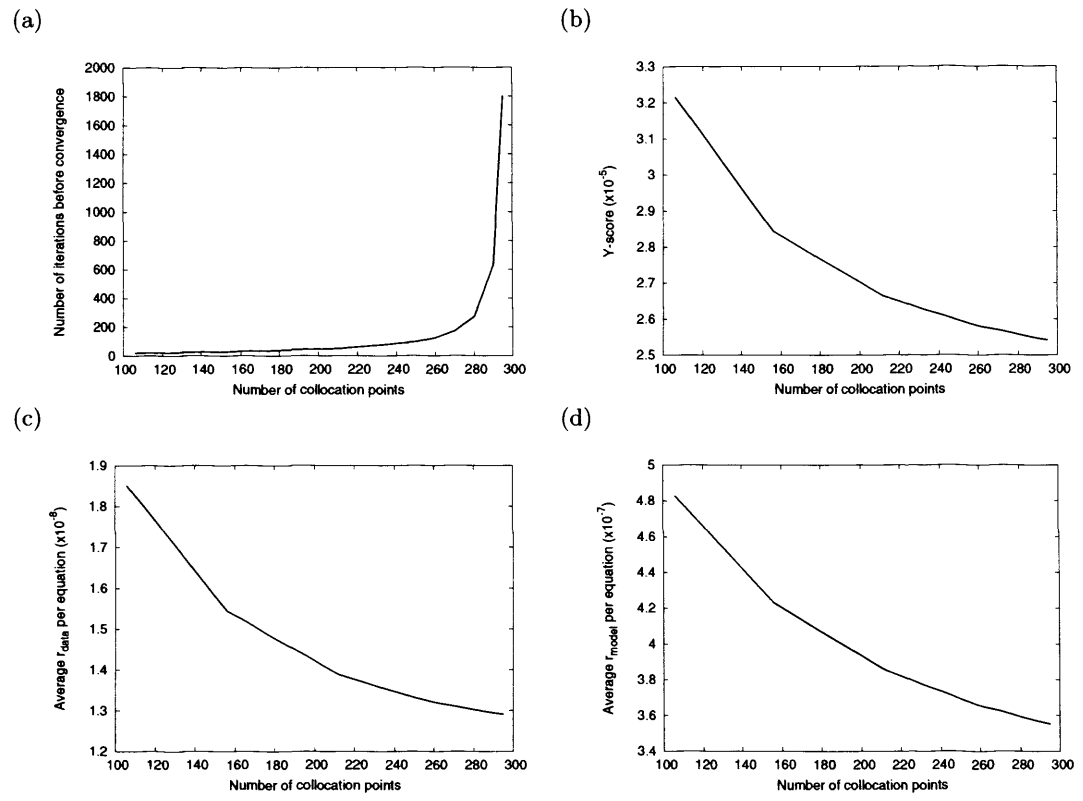


Figure 7.13: A plot showing how certain features vary as the number of collocation points is increased, (a) number of iterations before convergence, (b) the Y-score, (c) the average residual per equation of the data-spline equations and (d) the average residual per equation of the model-spline equations. This is based on the example system with 106 data points and 22 B-splines. No error was added to the data.

is the Y-score whose improvement supports the assumption that putting more weight on the model-spline equations will improve the estimated spline's fit to the true model solution. The average residual per equation is higher for the model-spline equations than the data-spline equations, this occurs because there is a limited number of B-splines and so the spline cannot satisfy the model exactly whereas the data is exact.

To get a better insight into the effect that the number of collocation points has on the performance of the algorithm, 1000 runs were performed on the 106 time point data set with 0.06 added error. This was repeated for a range of different numbers of collocation points from 106 to 312. As the number of collocation points is increased the number of convergent runs rapidly decreases at an increasing rate of decline (Figure 7.14(a)). There is no set point where the algorithm changes from being convergent to divergent, suggesting that the position of the data determines whether the algorithm is convergent. Only runs that were convergent are included in the rest of the results.

As the number of collocation points increases the average Y-score (see equation 7.8) decreases, appearing to tend towards some limit (Figure 7.14(b)). This is the same behaviour as is seen when there is no error (Figure 7.13(b)) and confirms that increasing

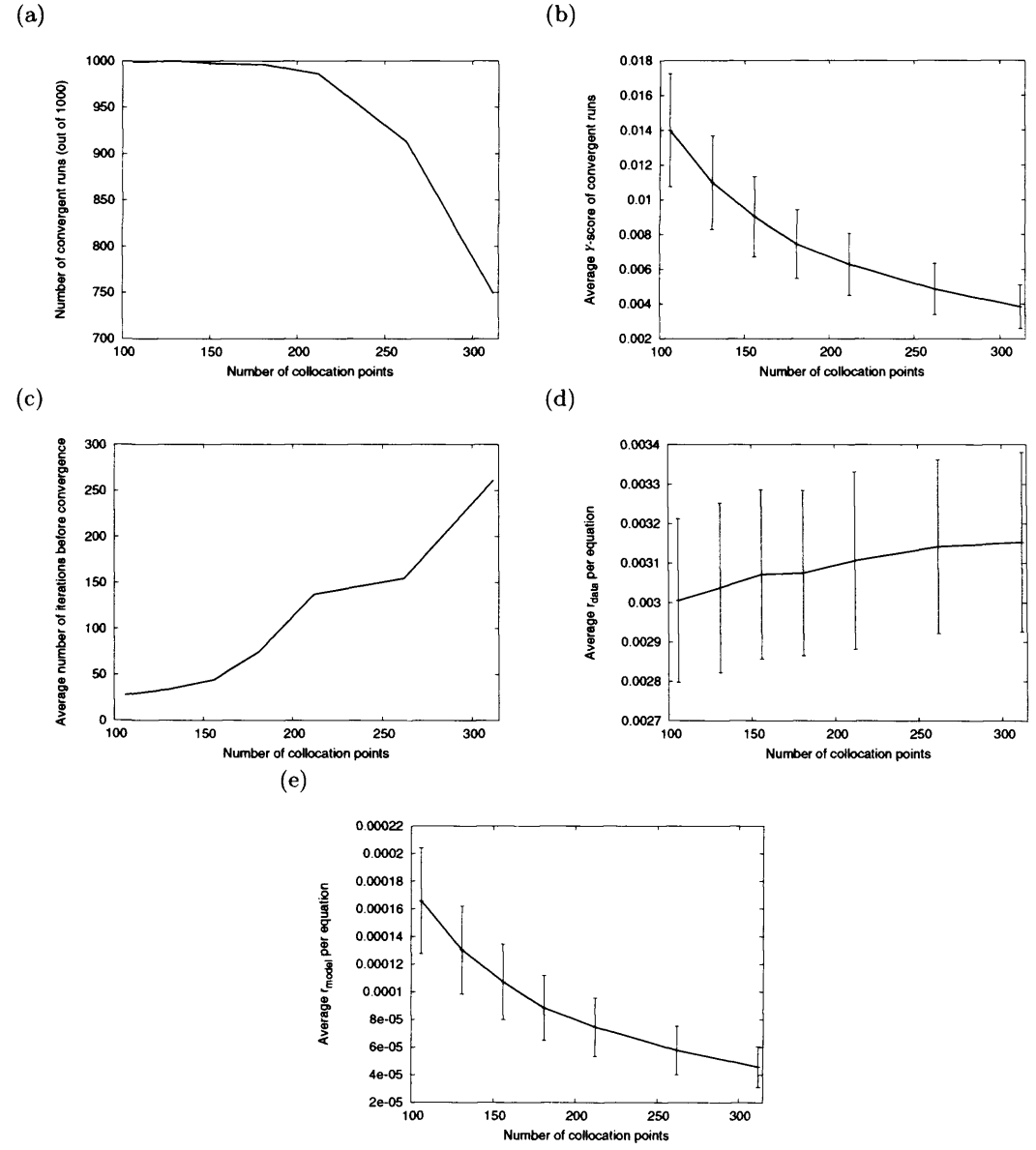


Figure 7.14: A plot showing how the algorithm performed when 0.06 error was added to the example data set with 106 time points on a variety of numbers of collocation points. There were 1000 runs but only results from those that converged are displayed. 22 B-splines were used. The following is shown, (a) the number of runs that converged, (b) the average Y-score, (c) the average number of iterations before convergence, (d) the average of the average r_{data} per equation and (e) the average of the average r_{model} per equation. The error bars, where shown, show the standard deviation of the quantity.

the spread and weight of the model-spline equations does produce better agreement between the solution spline and the model. The amount of improvement is much more significant when there is error, but the Y-scores are significantly higher. The improvement in the Y-score comes at a cost; on average the number of iterations before convergence increases as the number of collocation points is increased (Figure 7.14(c)). A possible explanation for this is that as the number of collocation points is increased the "pull" of the data points decreases, so it takes longer to draw the spline to its ideal position.

Interestingly, the average r_{data} per equation increases (Figure 7.14(d)), whereas it decreases when there is no error. The average r_{model} per equation decreases as the number of collocation points increases with and without error (Figure 7.14(e)). As n_c increases the improvement in r_{model} is much greater than the increase of r_{data}; a small worsening of the fit to the data greatly improves the spline fit to the model.

7.6.4 Re-weighing the equations

Increasing the number of collocation points improves the agreement between the solution spline and the model, but if there are too many collocation points the algorithm becomes divergent producing unsatisfactory results. This improvement was motivated by placing more weight on the spline agreeing with the model than with the data, by increasing the number of equations associated with the spline and model. An alternative would be to re-weigh the equations by multiplying through each row of the matrix equation by an appropriate factor. Here, only the model-spline equations (equation 7.2) will be re-weighed and they will be re-weighed by a factor,

$$\omega \times \frac{n_t}{n_c},$$

where ω is the *weight factor*. Multiplying the model-spline rows by $\frac{n_t}{n_c}$ has the effect of making the weight independent of the number of equations in each set and therefore the number of collocation points[6]. ω controls the relative weight between the two sets of equations and can be varied to get the best agreement between the spline and the model solution. The altered algorithm is as follows:

Algorithm 4. *An iterative method of parameter estimation with weight. This is the same as Algorithm 3 except that instead of equation 7.7, at each of the n_c collocation points, $t_0^c \ldots t_p^c \ldots t_{n_c-1}^c$, the variables should satisfy the following equation*

$$\omega \times \frac{n_t}{n_c} \sum_{j=0}^{n_s-1} \mathbf{b}_j B_j'(t_p^c) = \omega \times \frac{n_t}{n_c} \mathbf{f}\left(\mathbf{E}(t_p^c), \gamma\right), \tag{7.9}$$

where $\mathbf{f}()$ is a vector of functions linear in their parameters.

[6]It is advantageous to have a large number of collocation points as this spreads the positions where the model needs to be satisfied. This is important when there is a small amount of data.

Table 7.6: The parameter estimates from the algorithm with and without the reweighing when there is 500 collocation points. There are 22 B-splines and a data set containing 106 time points per component.

	D_{ATM}	D_{MDM2}	D_{p53}	p_{MDM2}	p_{p53}	k_1	k_2	k_3	k_4
True	0.05	0.18	0.041	0.01	0.52	1.42	0.39	2.50	0.75
With reweigh	0.05	0.18	0.0405	0.01	0.519	1.42	0.39	2.50	0.75
Without reweigh	0.05	0.275	−0.135	0.0540	0.363	1.29	0.268	2.70	0.807

This re-weighing is analogous with the general linear least squares optimisation (Press *et al.*, 2002), where each row of the matrix to be solved is associated with a data point. Each value in a particular row is divided by the standard deviation of the error distribution associated with that row's data point. In this way the general least squares puts more weight on the solution being close to data points with a small error. Similarly, having $\omega > 1$ is equivalent to saying that the model-spline equations are known to have less error associated with them than the data-spline equations.

7.6.5 Exploring the re-weighing

All experiments in this section were performed with 500 collocation points unless otherwise stated. With $\omega = 1$ the algorithm was run using the test model and the data set with 106 time points and no error. The algorithm converged after 10 iterations showing that this approach removes the control of convergence and weight away from the number of collocation points. The Y-score (2.357×10^{-5}) is slightly better than the Y-score found for Algorithm 2 when $n_c = 106$ (3.22×10^{-5}). The Y-score is similar to what would be found if Figure 7.13 was extrapolated to 500 collocation points. The parameters estimated are very good, much better than if there was no reweighing (Table 7.6). The algorithm still becomes divergent when the weight is increased too far, in this case the transition occurred when $\omega \approx 4$.

An example random data set was produced using 0.06 error. The algorithm was found to become divergent on this data set at $\omega = 3.4$. The results when ω is varied (Figure 7.15) are similar to those observed previously (Figure 7.14); the Y-score decreases with increasing weight and the number of iterations before convergence increases. When $\omega = 3.3$ (close to the optimal value) the solution spline produced by Algorithm 4 is smoother with less peaks and troughs than Algorithm 2 (the original iterative algorithm) (Figure 7.16). The Algorithm 4 spline is slightly closer to the true solution than the spline produced by Algorithm 2, but both spline's follow the same gross dynamics which are not that close to the true solution. In this situation though, the "true" solution may not be close to the least squares model solution because of the large amount of error in the data. A better indicator of the quality of the spline is the Y-score, which seems reasonable for Algorithm 4 ($Y = 0.00352$).

To gain a more generalised view on what happens with regard to the weight and

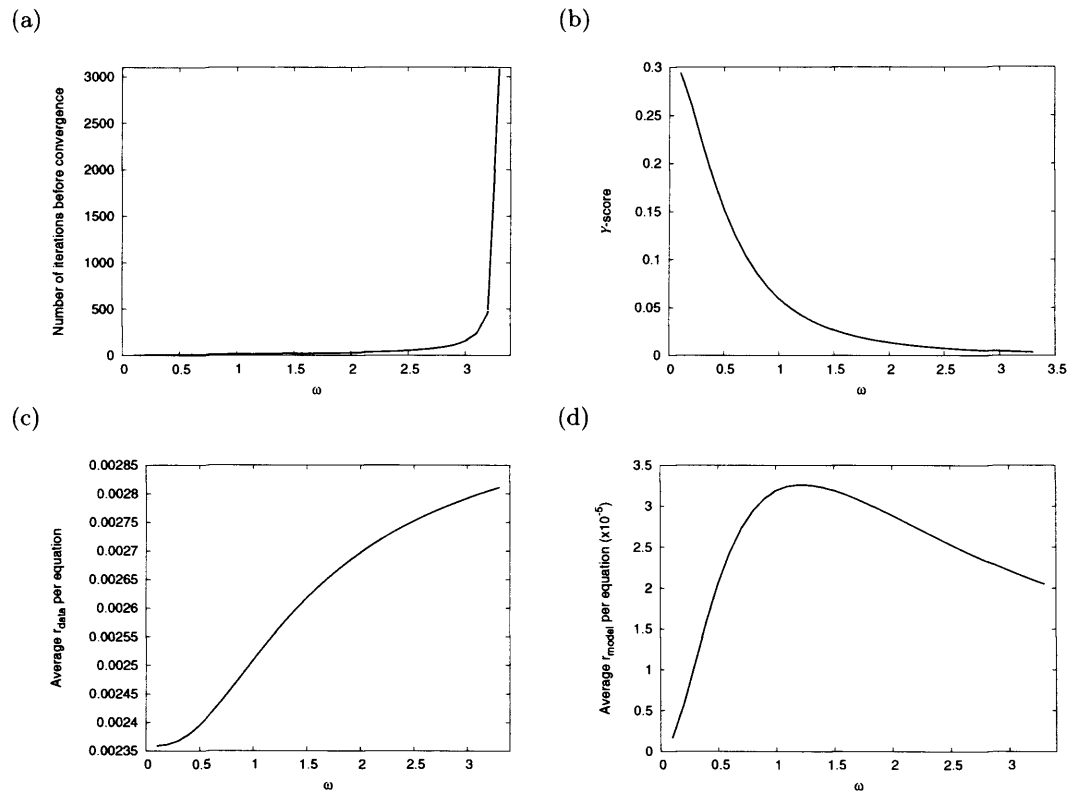


Figure 7.15: A plot showing how certain features vary as ω is increased on a data set with 0.06 error, (a) number of iterations before convergence, (b) the Y-score, (c) the average residual per equation of the data-spline equations and (d) the average residual per equation of the model-spline equations. This is based on the example system with 106 data points, 500 collocation points and 22 B-splines. The equivalent values for the non-adapted algorithm (Algorithm 2) are: iterations = 36, Y-score = 0.0110, average residual per equation for the data-spline set is 0.00272 and for the model-spline set is 0.000130.

noisy data the algorithm was run 1000 times, each time with a new 0.06 error added data set. The algorithm's performance (Figure 7.17) was similar to that observed when the number of collocation points were used as the weight control rather than ω (Figure 7.14). Interestingly, the Y-score (Figure 7.17(b)) starts higher and ends lower than for Algorithm 3 (Figure 7.14(b)), indicating that the range of ω produces a larger range of weights on the equation sets than the number of collocation points. Also, the Y-score might be converging to a lower value because 500 collocation points were used here. The data-spline residual (Figure 7.17(d)) has similar behaviour to previously and is at approximately the same values (Figure 7.17(d)). The model-spline residual (Figure 7.17(e)) has smaller values than previously (Figure 7.14(e)), probably because there is a greater number of collocation points.

The distribution of the estimated parameters do not have a common behaviour as ω increases (Figure 7.18), some parameter values increase, some decrease, and others peak. The average parameter value does not approach the true parameter value when ω

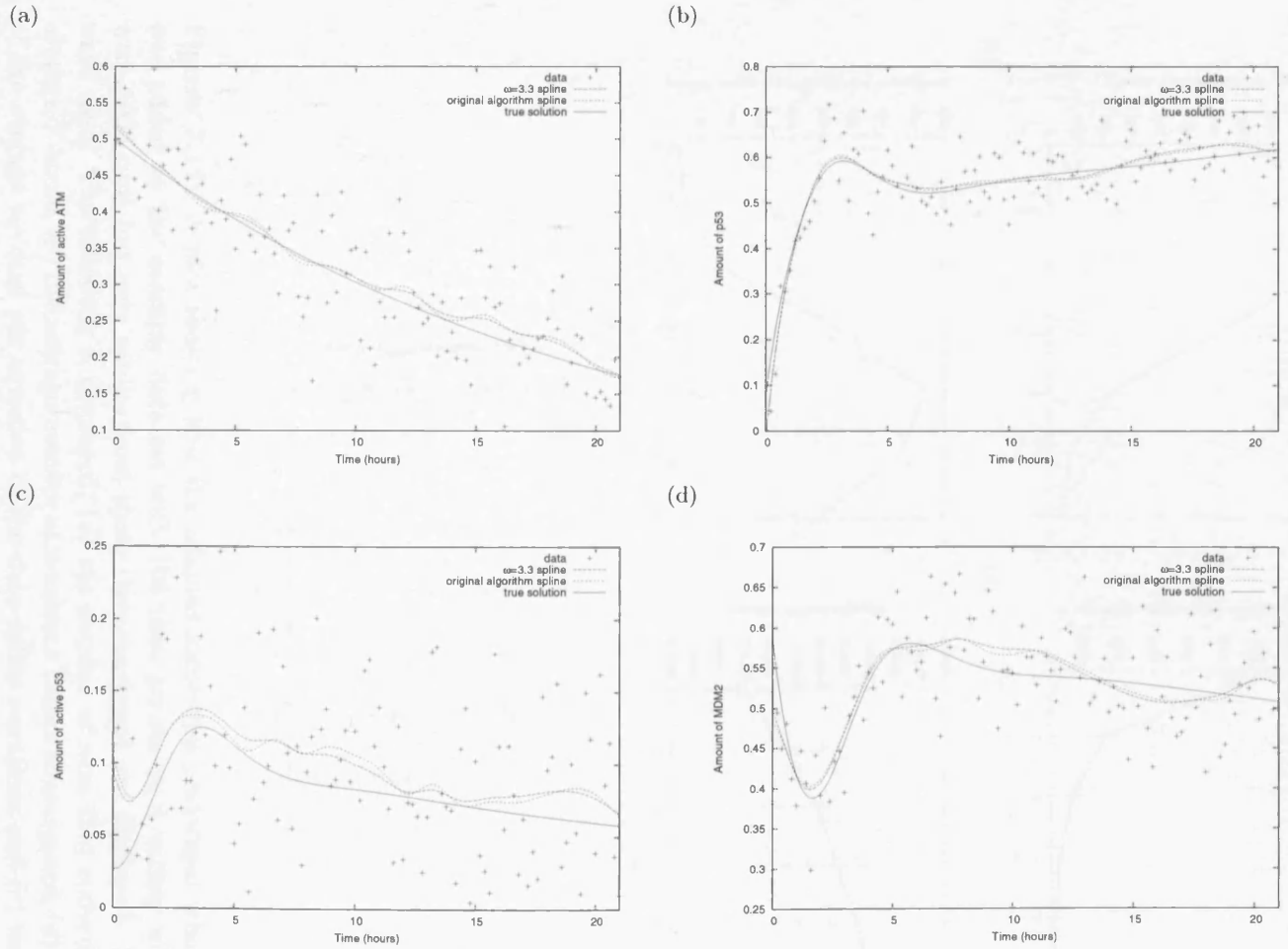


Figure 7.16: A plot comparing the solution splines of Algorithm 2 and Algorithm 4 with $\omega = 3.3$ (close to the limit) for (a) active ATM, (b) p53, (c) active p53 and (d) MDM2. This is based on the example system with 106 data points (0.06 added error), 500 collocation points and 22 B-splines.

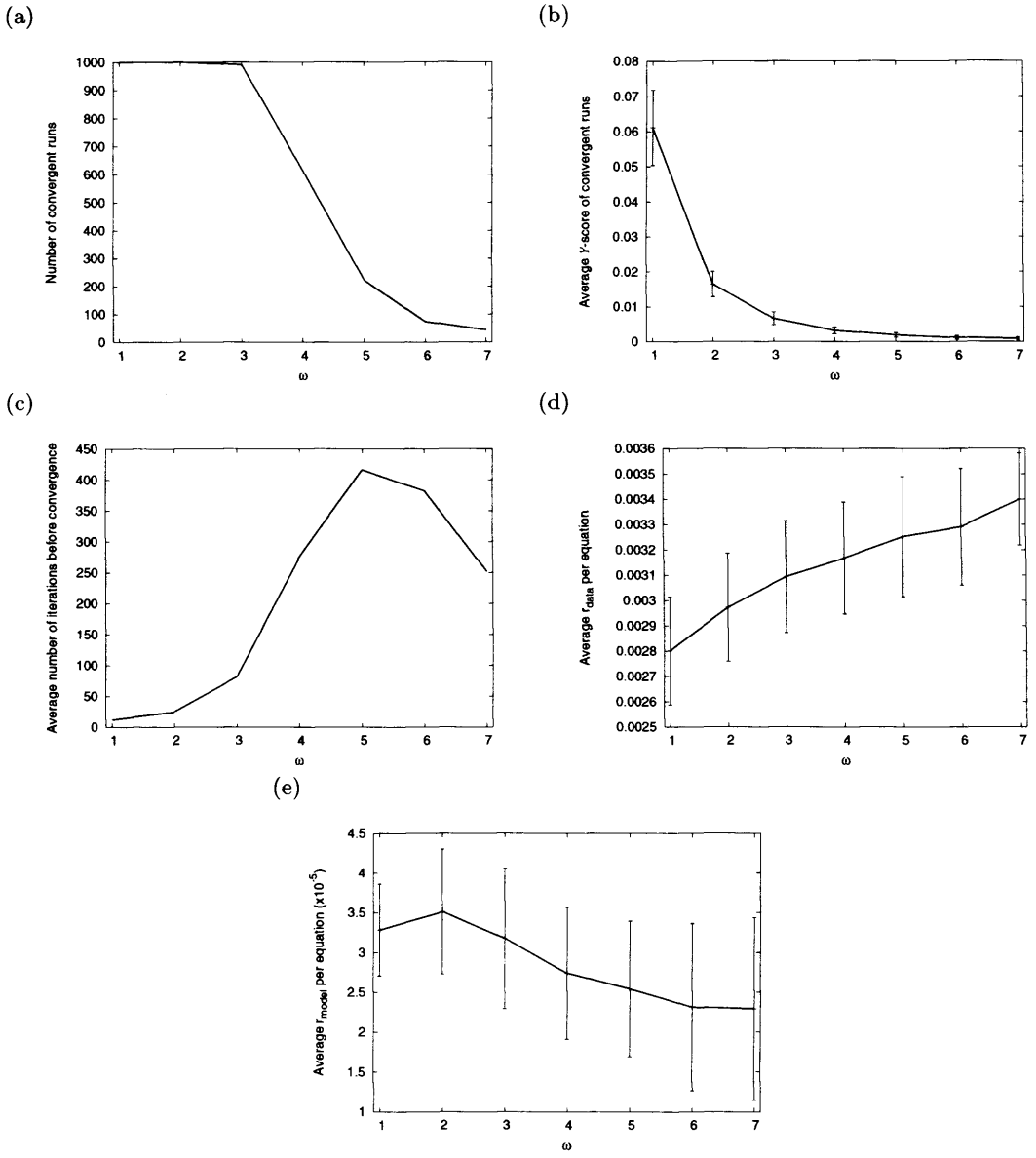


Figure 7.17: A plot showing how the adapted algorithm performed when 0.06 error was added to the example data set with 106 time points on a variety of ωs. There were 1000 runs but only results from those that converged are displayed. 22 B-splines were used. The following is displayed, (a) the number of runs that converged, (b) the average Y-score, (c) the average number of iterations before convergence, (d) the average of the average residual per equation of the data-spline equations and (e) the average of the average residual per equation of the model-spline equations. The error bars, where shown, show the standard deviation of the quantity.

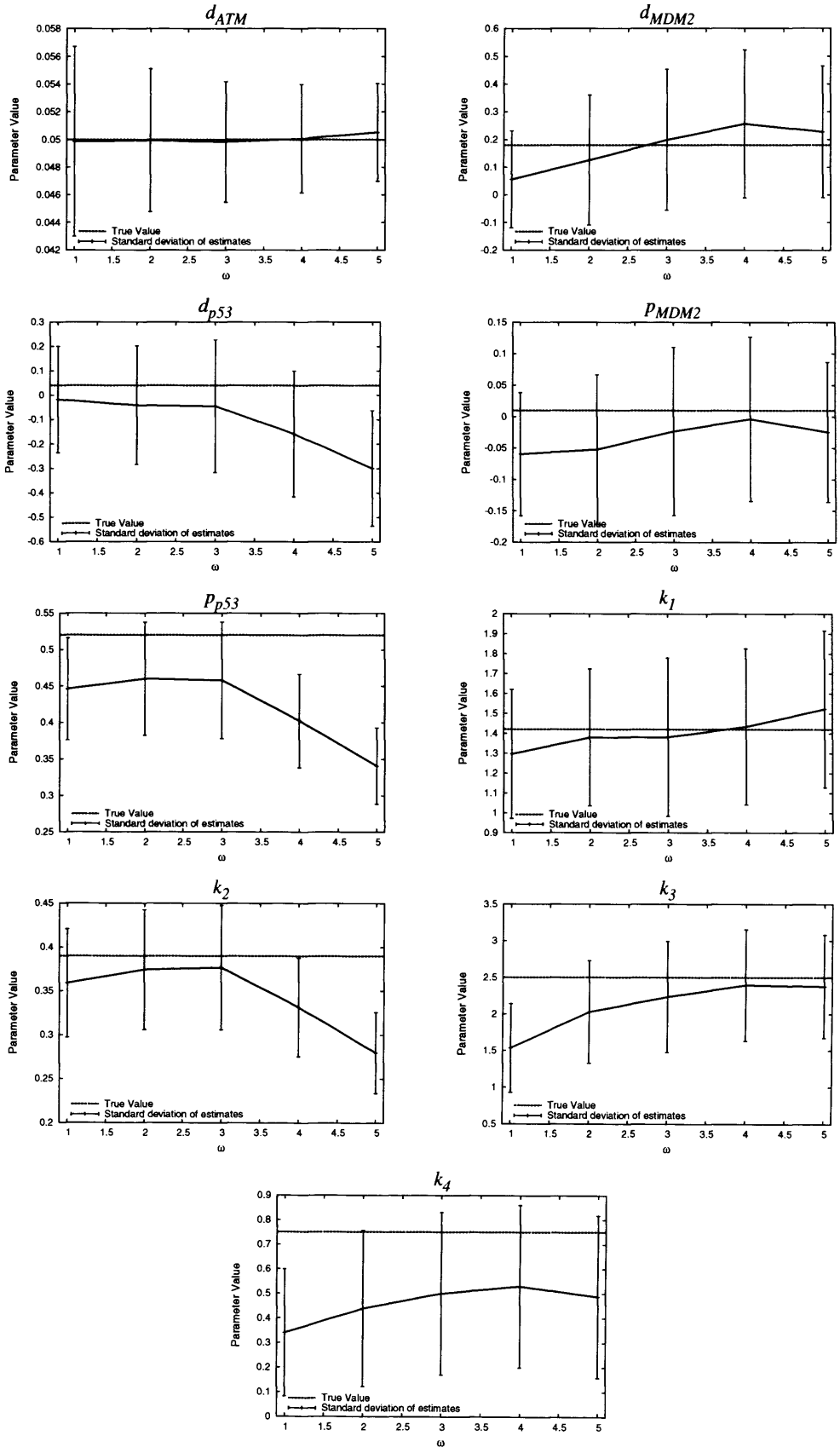


Figure 7.18: A plot of how the parameter estimates change with ω for Algorithm 4. 0.06 error was added to the example data set with 106 time points. Based on 1000 runs but only results from those that converged are displayed. 22 B-splines were used.

is large. This suggests that increasing ω does not necessarily improve the quality of the parameter estimates, but this can be discounted because:

1. There is a large amount of error in the data

   As there is a large amount of error the data points will diverge significantly from the true values. This means that there is a considerable likelihood that that the net behaviour will be different from the true data set. Therefore the smallest distance measure by any optimisation method will not produce a model solution that will be close to the "true" model solution and so this means that the best parameter values for a particular data set is not the same as the true parameter values.

2. Only the convergent runs have been included in the analysis

   Some selection is occurring because only convergent runs have been considered. Even if there is a large amount of error as described above one would still expect that the average model behaviour would be close to the true model behaviour. In this experiment though, excluding the divergent data is effectively acting as a filter. It is assumed that the closer the best model solution is to the true model solution then the closer the transition ω is to the no error case. This has the effect that those data sets whose optimal parameter values are close to the "true" parameter values will be cut out at a similar weight as the no error case. When there was no error the transition occurred at $\omega \approx 4$, and it is generally true that for $\omega > 4$ the parameter values move away from the true value.

Depending on the data set the optimal ω may result in a Y-score that is still high. In this situation a satisfactory parameter estimate may not be produced by this algorithm.

7.7 Low amounts of data

The amount of experimental data available will generally be small and this produces problems for any parameter estimation method, but there are some particular problems for the algorithms proposed here. As observed in section 7.3.1, as the number of data points is decreased the parameter estimates get increasingly worse, even for data with no error. There are a number of reasons that this occurs even though the residuals are small:

1. Even though the spline satisfies the model at the data points, there is no constraint between the data points. This means that across the whole of the spline it is considerably different from the model solution. If the spline does not represent the model solution well then the parameter estimates will be poor.

2. The number of data points restricts the number of B-splines that can be used. As the number of B-splines decreases, the spline's capacity to accurately resemble the model solution is diminished.

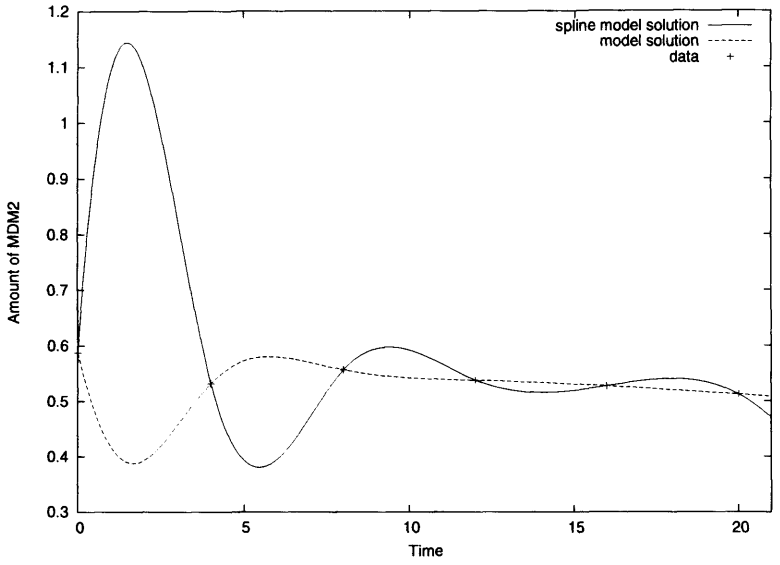


Figure 7.19: An example of the spline produced when there is a low amount of data. This is using the simple model and the trace of MDM2 is shown. There are 6 data points and 8 B-splines in the spline.

Table 7.7: The parameter estimates when there are 8 B-splines and a data set containing 6 time points per component.

	D_{ATM}	D_{MDM2}	D_{p53}	p_{MDM2}	p_{p53}	k_1	k_2	k_3	k_4
True	0.05	0.18	0.041	0.01	0.52	1.42	0.39	2.50	0.75
Estimated	0.05	−2.63	−0.0468	−1.05	0.581	1.75	0.460	−8.58	−2.07

An example of this is when Algorithm 2 is run with 8 B-splines on a data set of 6 time points with no error (Figure 7.19). The residuals are low at $r_{\text{data}} = 1.68 \times 10^{-6}$ with at least half the parameter estimates being poor (Table 7.7). A low amount of data also makes the algorithm sensitive to the other problems that occur when there is error in the data.

7.7.1 Increasing the number of collocation points

At a low number of data points it was possible to get a small residual in the model-spline set of equations but the fit of the solution spline to the model was very poor i.e. the Y-score was high (see equation 7.8). This is because the spline only has to satisfy the model at the data points. If the number of collocation points are increased, the spline will attempt to satisfy the model at a far greater range of points and so the spline is likely to be closer to the model solution. In this section the effect of increasing the number of collocation points and the use of Algorithm 4 on a small data set will be examined.

Using the example model (see section 4.3.3), the algorithm was run on a data set with 6 time points and 8 B-splines (the maximum possible). Algorithm 2 gave a Y-score of 16.2, the same score as Algorithm 4 with 6 collocation points and $\omega = 1$. When ω

Table 7.8: The parameter estimates in a number of different situations. There are 8 B-splines and a data set containing 6 time points per component.

n_c	D_{ATM}	D_{MDM2}	D_{p53}	p_{MDM2}	p_{p53}	k_1	k_2	k_3	k_4
True Values	0.05	0.18	0.041	0.01	0.52	1.42	0.39	2.50	0.75
Pre-adapted algorithm	0.05	−2.63	−0.0468	−1.05	0.581	1.75	0.460	−8.58	−2.07
17 collocation points $\omega = 2.743$	0.05	−0.0938	0.874	−0.0638	0.497	−0.225	0.369	0.339	0.104

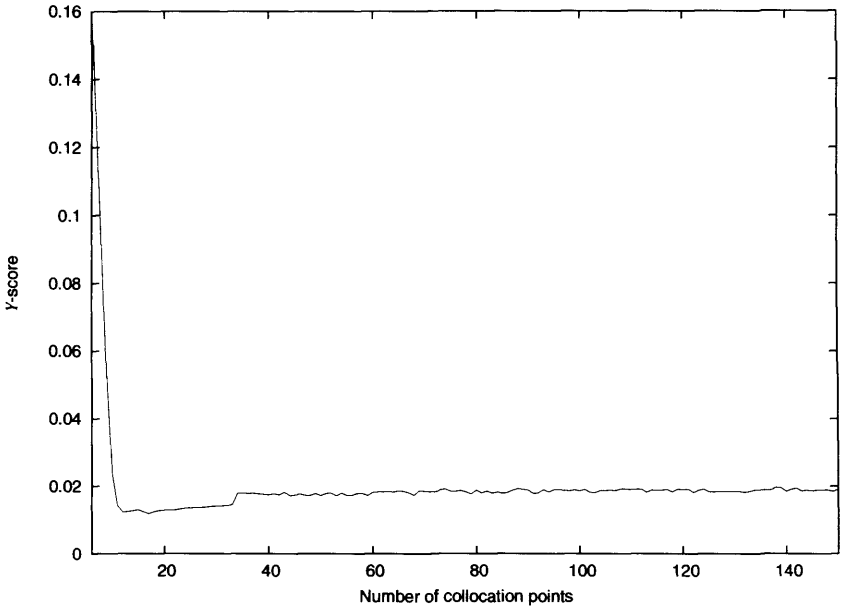
is increased to 2.4 a dramatic improvement is made with a Y-score of 0.162, confirming that increasing the weight placed on the spline agreeing with the model, improves the results of the algorithm.

Algorithm 4 was run on a wide range of n_cs with ω set to the maximum possible value in each instant (to a resolution of 0.001). As the number of collocation points was increased it was expected that that the Y-score would improve. Initially, this was the case with Y dropping rapidly until approximately $n_c = 17$, with an order of magnitude improvement since the initial score (Figure 7.20(a)). After this Y increases slowly until $n_c = 34$ where it jumps by 0.003 and then increases at a slow rate. It is unclear why this should occur, but it could be to do with an increasing number of collocation points producing greater numerical errors, or a subtle ω problem. As the number of collocation points increases, the maximum ω that produces convergent behaviour also increases (Figure 7.20(b)). These results suggest that a search over a range of collocation points must be made to get the best possible results. When the collocation number was set to 7 or 8 no ω could be found where the algorithm was convergent. This shows that it is not always possible to use this parameter estimation technique with low data sets.

When there are 17 collocation points the optimal Y-score of 0.0119 is achieved (with a maximum ω of 2.743). By increasing the number of collocation points and increasing the weighting, the Y-score has been improved by over 1000 fold which is an improvement. Increasing the number of collocation points and weight factor also dramatically improves the parameter estimates (see Table 7.8). The parameters are still not close to the true values but are reasonable considering the number of data points. In particular the estimates for D_{ATM}, p_{p53} and k_2 are good.

The solution splines produced by Algorithm 4 when there are 500 collocation points and Algorithm 2 are reasonably similar for both active p53 and p53 (Figure 7.21). For MDM2 though, there is a significant change; in the first 4 hours the Algorithm 2 spline has a peak whereas the Algorithm 4 spline has a trough, which agrees with the true solution (the amplitude of the trough though is not as deep as the true solution). Generally the fit of the splines is not accurate to the true solution because of the small amount of data and the restriction on the number of B-splines.

(a)



(b)

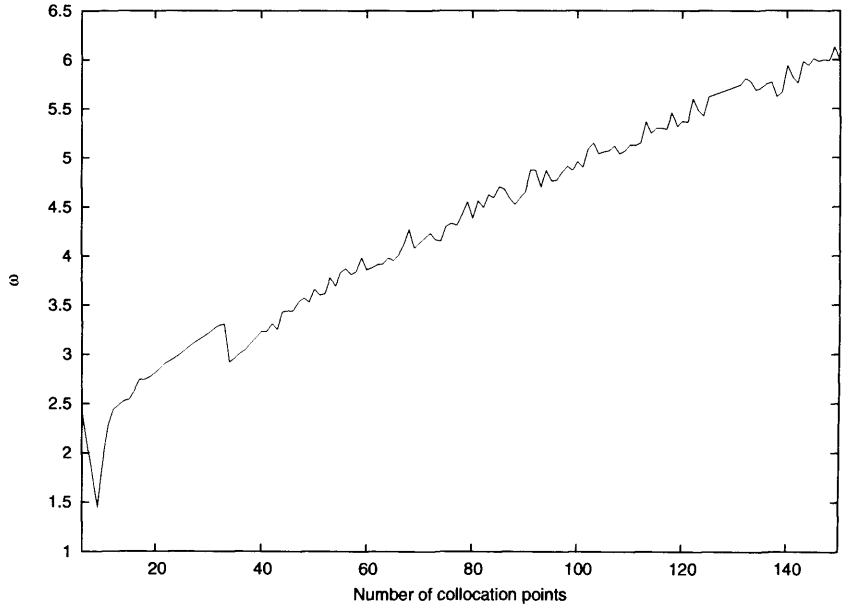
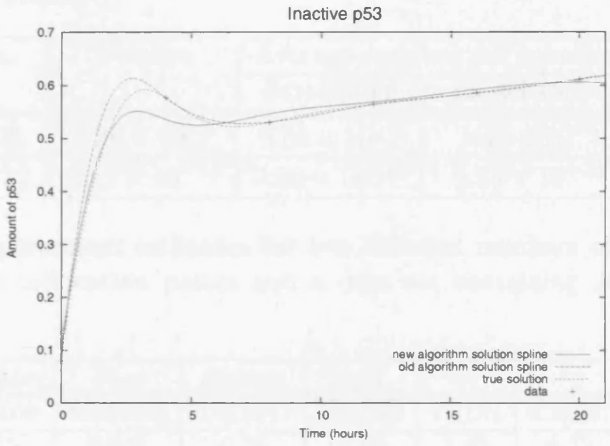
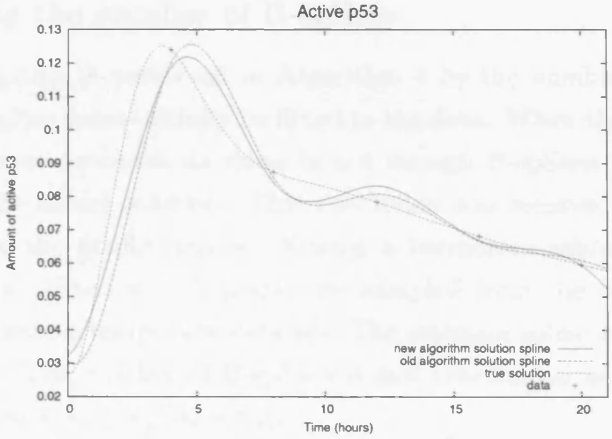


Figure 7.20: A plot of how (a) the Y-score (b) maximum ω varies with the number of collocation points used in Algorithm 4. This was performed on a 6 time point data set with 8 B-splines. The maximum ω possible was used. It only includes data which had a stable solution.

(a)



(b)



(c)

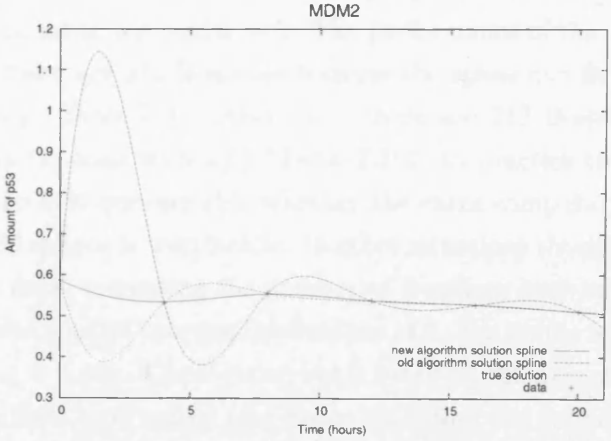


Figure 7.21: A plot comparing the solution splines of the Algorithm 2 and the adapted Algorithm 4 with 500 collocation points and $\omega = 10$ (close to the limit). This is based on the example system with 6 data points and 8 B-splines. ATM is not shown as both solution splines fit the true model solution extremely well.

Table 7.9: The results on the 106 time point data set with 22 and 212 B-splines. There are 500 collocation points.

n_s	Y-score	Average residual per equation	
		data-spline	model-spline
22	2.36×10^{-5}	8.98×10^{-9}	1.43×10^{-8}
212	2.52×10^{-11}	7.96×10^{-18}	1.33×10^{-14}

Table 7.10: The parameter estimates for two different numbers of B-splines. This is for a run with 500 collocation points and a data set containing 106 data points per component.

n_s	D_{ATM}	D_{MDM2}	D_{p53}	p_{MDM2}	p_{p53}	k_1	k_2	k_3	k_4
22	0.05	0.18009	0.040498	0.010047	0.51946	1.4194	0.38957	2.5002	0.75012
212	0.05	0.18	0.041	0.01	0.52	1.42	0.39	2.5	0.75

7.7.2 Increasing the number of B-splines

The number of B-splines is restricted in Algorithm 4 by the number of time points to $n_t + 2$ because the spline must initially be fitted to the data. When there is a low amount of data this will cause problems as there is not enough B-splines to produce a good approximation of the model solution. This restriction was removed by introducing an intermediate step in the fitting routine. Firstly, a temporary spline with n_t B-splines is fitted to the data. Then $n_s - 2$ points are sampled from the temporary spline at equal intervals to create a temporary data set. The estimate spline is then fitted to this temporary data set. The number of B-splines is still constrained as the problem needs to be over-defined ($n_t + n_c \geq n_p/n_v + n_s$).

To ensure that this adaptation is working as expected it was initially run on the 106 time point data set for two different amounts of B-splines: 22 and 212. The number of collocation points was set to 500 and $\omega = 1$. The performance of the algorithm is considerably better when there are 212 B-splines because the spline can fit the model solution with greater accuracy (Table 7.9). Also when there are 212 B-splines the parameter estimates are perfect (at least to 5 *s.f.*) (Table 7.10). In practice this level of accuracy is not required and so it is questionable whether the extra computational time required when there is more B-splines is worthwhile. In other situations though, such as when are fewer or more noisy data, increasing the number of B-splines may be productive.

This algorithm was applied to a smaller data set of 6 time points with 22 B-splines per spline, $n_c = 29$ and $\omega = 0.464$. The Y-score was 5.3×10^{-5}, the parameter estimates were remarkably close to their true values (see Table 7.11) and the spline solution was close to the true solution (see Figure 7.22 for an example). When there are 14 B-splines and 14 collocation points the parameter estimates were an improvement over the 8 B-spline case but not as good as the 22 B-spline case (Table 7.11). The Y-score was 6.98×10^{-4} and the resulting splines were accurate but not quite as good as the 22 B-spline situation

Table 7.11: The parameter estimates for the algorithm applied to a data set containing 6 data points per component.

n_c	D_{ATM}	D_{MDM2}	D_{p53}	p_{MDM2}	p_{p53}	k_1	k_2	k_3	k_4
True Values	0.05	0.18	0.041	0.01	0.52	1.42	0.39	2.50	0.75
8 B-splines $n_c = 17$ $\omega = 2.743$	0.0500	−0.0938	0.874	−0.0638	0.497	−0.225	0.369	0.339	0.104
22 B-splines $n_c = 29$ $\omega = 0.464$	0.0500	0.176	0.0391	0.0954	0.525	1.44	0.394	2.47	0.741
14 B-splines $n_c = 14$ $\omega = 0.464$	0.0500	0.212	0.0391	−0.00914	0.406	0.984	0.307	3.68	1.13

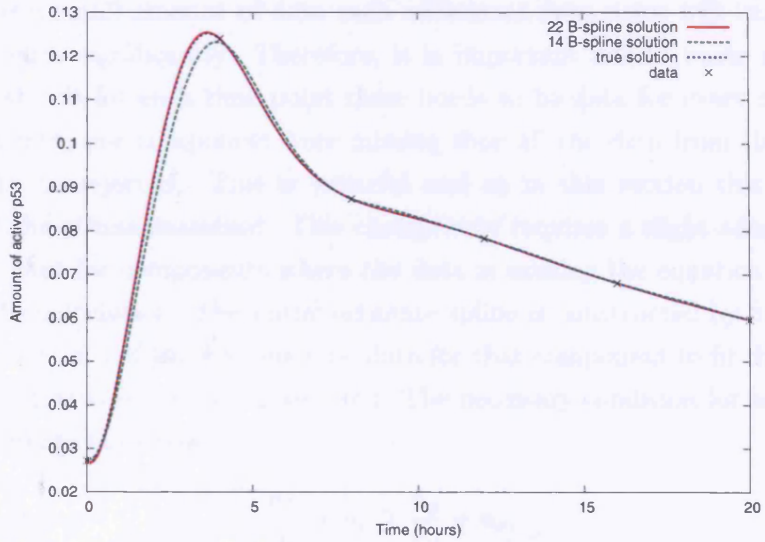


Figure 7.22: The solution spline for active p53 produced by the adapted Algorithm 4 when run on a data set of 6 time points when there are (a) 22 B-splines, 29 collocation points & $\omega = 0.464$ and (b) 14 B-splines, 14 collocation points & $\omega = 1.09$.

(Figure 7.22). Increasing the number of B-splines can significantly improve the parameter estimates when there is a small amount of data. It is unlikely that estimates as good as when there are 22 B-splines would be achievable if there was error in the data. Also, introducing more B-splines adds variables to the problem and could make it problematic to get the solution spline to be smooth.

The above experiments on the low data set were hampered by the algorithm diverging, it was very difficult to find values of ω and n_c so that the algorithm would converge. For many combinations of B-spline number and n_c there was no convergence at all, whatever the ω value (see Table 7.12). The most likely explanation for this tendency to diverge is that increasing the number of B-splines also increases the number of variables making the problem more difficult to solve; there is not enough data to anchor and hence stabilise the spline when there are a high number of variables.

Table 7.12: The combination of n_c and number of B-splines (n_s) that give convergent runs, based on a search of n_c between 6 & 30 and an even n_s between 10 & 22.

Number of B-splines	n_c	Max. ω	Y-score
10	8	19.9	0.242
12	24	0.1	0.0238
14	14	1.09	6.98×10^{-4}
14	15	1.11	0.00313
22	23	0.8	7.67×10^{-5}
22	29	0.464	5.3×10^{-5}

7.7.3 Fixing for when data is missing

When there is a small amount of data each additional data point will improve the parameter estimates significantly. Therefore, it is important not to waste any data. As Algorithm 4 stands for each time point there needs to be data for every model component. If data from one component were missing then all the data from that time point would have to be rejected. This is wasteful and so in this section this restriction is loosened and the effects examined. This change only requires a slight adaptation of the algorithm, in that for components where the data is missing the equation is simply not included in the calculation. The initial estimate spline is constructed by fitting it to the remainder of the data, if there is too little data for that component to fit the spline, then the steps described in section 7.7.2 are used. The necessary condition for an over-defined set of linear equations is now,

$$\frac{n_D}{n_v} + n_c \geq \frac{n_p}{n_v} + n_s,$$

where n_D is the total number of data points (when there is no data missing $n_D = n_t n_v$).

Various amounts of data were removed from the 106 time point data set and then used with the adapted algorithm. This was repeated 1000 times. The data removed was randomly chosen and an equal amount of data was removed from each component. The parameter estimates remain very good even when 70 out of 106 data points have been removed (Figure 7.23). The averages of the estimates generally move away from their original values but only by a small amount. As the number of data points removed is increased the most significant change is that the distribution widens. This shows that the precise data points removed can have a significant effect on the parameter estimates. When a large amount of data is removed, inaccurate parameter estimates may be caused by a large gap in an important part of the systems profile.

The algorithm can fail when the initial spline is fitted to the data, and the proportion of failed runs increases as the amount of data removed is increased (Figure 7.24). This occurs when there are large holes in the data, bigger than the range of a B-spline, making it impossible to fit that B-spline to the data. This occurs more at the beginning or the

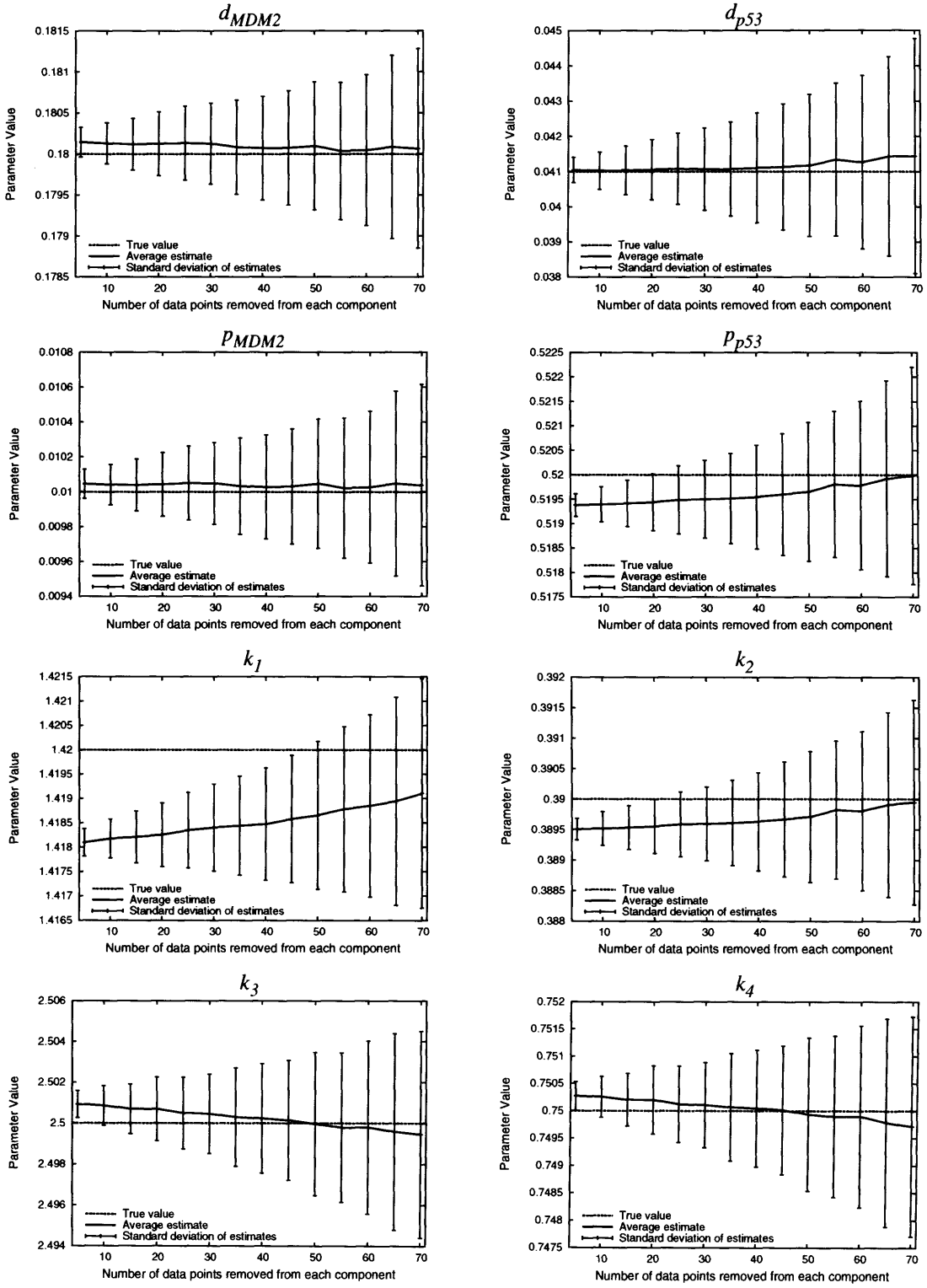


Figure 7.23: A plot showing how the parameter estimation changes when an increasing number of time points are removed from the 106 data set. The number of B-splines was set to 22, $n_c = 106$ and $\omega = 1$. D_{ATM} is not shown as the results were very accurate whatever the amount of data removed

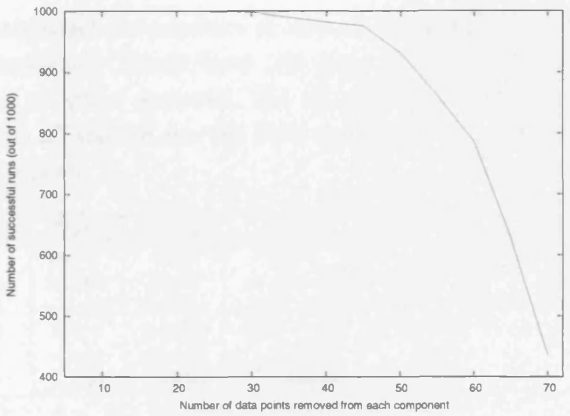


Figure 7.24: A plot showing how the number of successful runs decreases as the points removed increases. The number of B-splines is set to 22, $n_c = 106$ and $\omega = 1$.

end of the data due to the "free" B-splines at the extremes. This restricts the data that can be used with the algorithm, but could be overcome by interpolating data points in the large gaps for the initial fit.

In a biological setting a particular component of the model might be difficult, if not impossible, to measure in an accurate way. Is it possible to use the algorithm to gain information about the parameters even when there is no data for a particular component? A particular problem of this situation is that an initial estimate spline has to be constructed without any knowledge about where it should go. One approach to overcome this is to set the spline as a straight line at an arbitrary value, in this case at a level of one unit. It would be beneficial if one data point could be found for the missing component so that the level could be set at a reasonable value. To test this on the example 106 data set, the data for inactive p53 was removed. The parameter estimation routine was not convergent when there was no data points or when there was just one data point because there was nothing to anchor the B-spline. Despite this the majority of the parameter estimates are surprisingly good (see Table 7.13(a)-(c)). The parameters that are not close to their true values are p_{p53}, k_1 and k_2 which are all involved with the level of p53[7].

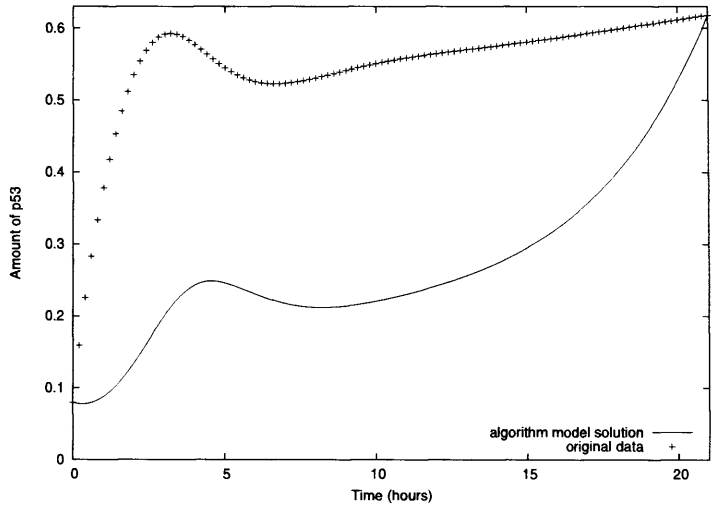
When there are two data points for p53, the situation is improved further, with all the parameter estimates at the right order of magnitude (see Table 7.13(d) and (e)). In this case the routine has confused D_{p53} and p_{p53} with both having negative values. Even though the corresponding model solutions for p53 could not be described as accurate, the solutions do have similarities with the true solution; in both cases there is a peak in approximately the same place (Figure 7.25). The solutions for the other components are good fits. It is interesting that even when the second data point is fixed at a completely inaccurate value the estimates are restricted to reasonable values and there is some information about the model solution. This suggests that even the information that a

[7] k_1 is the rate that p53 is activated by ATM and k_2 is the rate that p53 & active p53 is degraded by MDM2.

Table 7.13: The estimated parameters produced when all or nearly all of the inactive p53 data has been removed. There were 106 time points, 22 B-splines, 106 collocation point and $\omega = 1$. (a) all data removed, (b) all apart from the first point, (c) all apart from the last point, (d) all except for the first and last and (e) all except for the first and the last point is set to zero.

	D_{ATM}	D_{MDM2}	D_{p53}	p_{MDM2}	p_{p53}	k_1	k_2	k_3	k_4
True	0.05	0.18	0.041	0.01	0.52	1.42	0.39	2.5	0.75
(a)	0.0500	0.218	0.0695	0.0237	-6.20×10^{9}	-2.26×10^{-10}	$1.39 \times 10{-}11$	2.70	0.809
(b)	0.0500	0.209	−0.120	0.0183	-2.97×10^{8}	-1.17×10^{-9}	4.74×10^{-10}	2.71	0.812
(c)	0.0500	0.193	0.167	0.0120	2.30×10^{8}	3.48×10^{-11}	-2.2×10^{-10}	2.63	0.790
(d)	0.0500	0.241	−1.71	0.0420	−0.0153	3.02	−0.121	2.54	0.766
(e)	0.0500	0.247	−1.29	0.0450	−0.000897	2.52	0.267	2.58	0.791

(a)



(b)

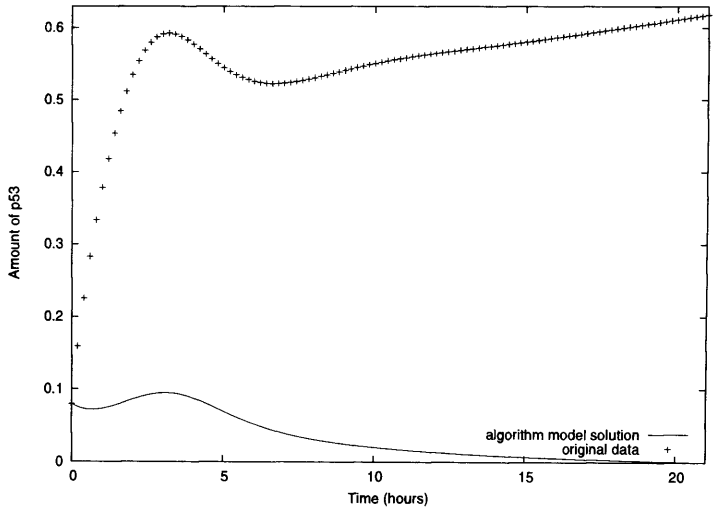


Figure 7.25: A plot showing the predicted model solution for p53 when there are only two p53 data points, (a) the first and last data point and (b) the first data point and the last data point set to 0.

Table 7.14: The parameter estimates with and without Nelder-Mead solving the set of equations. There are 22 B-splines, $n_c = 500$, $\omega = 1$ and a data set containing 106 time points per component.

	D_{ATM}	D_{MDM2}	D_{p53}	p_{MDM2}	p_{p53}	k_1	k_2	k_3	k_4
True Values	0.05	0.18	0.041	0.01	0.52	1.42	0.39	2.50	0.75
with out Nelder-Mead	0.05	0.180	0.0405	0.0100	0.519	1.42	0.390	2.50	0.750
with Nelder-Mead	0.05	0.180	0.0403	0.0101	0.519	1.42	0.390	2.50	0.750

component rises or falls during a response could be useful.

It is unwise to rely on the parameter estimation routine to produce reasonable values when there is no data for a model component. It has been shown though that even in this situation some information can be gained about the system. Adding additional experimental data, even a small amount, will significantly improve the performance of the parameter estimation routine.

7.7.4 Dealing with negative parameter values

A noticeable problem with the parameter estimation routine is that the parameter estimates are sometimes negative, they are not restricted to positive values. This occurs especially when there is error in the data or a low amount of data. Parameters that are particularly liable to become negative are the production and degradation terms of the various components, if both of these become negative then the effect can be approximately the same as both being positive (for example D_{MDM2} and p_{MDM2} in Table 7.8).

One way to restrict the parameters to positive values is to move the model parameters into the log domain. This means that the parameter estimation routine will work in parameters that cover both positive and negative numbers, but when these parameters are applied to the model the exponential of the parameters are used. This confines the actual parameters to the positive domain. Unfortunately this means that the equations are no longer linear in their parameters and so linear algebra cannot be used. Therefore the Nelder-Mead downhill simplex method will be used to solve the set of equations (Nelder and Mead, 1965), with the rest of the algorithm remaining the same. It is likely that there will still be one solution in the log-domain. One iteration is performed without using Nelder-Mead and the change to the log-domain so that the Nelder-Mead parameter estimation can be seeded with reasonable values. If the seed parameters are negative they are given a value of -100, which is effectively zero.

To test that the Nelder-Mead solver is working correctly within Algorithm 4, the algorithm was run on the 106 data set with 22 B-splines, $n_c = 500$ and $\omega = 1$. The parameter estimates are virtually the same as when Nelder-Mead was not used within the algorithm (Table 7.14). The resulting splines are a close fit to the real time course.

The algorithm with the Nelder-Mead solver was then applied to the small data set.

Table 7.15: The parameter estimates with and without Nelder-Mead solving the set of equations. There are 8 B-splines, $n_c = 17$, $\omega = 2.743$ and a data set containing 6 time points per component.

n_c	D_{ATM}	D_{MDM2}	D_{p53}	p_{MDM2}	p_{p53}	k_1	k_2	k_3	k_4
True	0.05	0.18	0.041	0.01	0.52	1.42	0.39	2.50	0.75
without simplex	0.05	−0.0938	0.874	−0.0638	0.497	−0.225	0.369	0.339	0.104
with simplex	0.0501	2.61×10^{-35}	3.63×10^{-59}	3.10×10^{-31}	0.505	1.44	0.372	0.146	0.0872

The parameter estimates are an improvement over those when Nelder-Mead was not used with one extra good estimate (k_1) and with most of the other good estimates being closer to their true value (Table 7.15). Three of the parameters are estimated as being approximately equal to zero (D_{MDM2}, D_{p53} and p_{MDM2}) and these are the parameters that were negative when the Nelder-Mead routine was initialised. Given that the basal rates associated with MDM2 are estimated incorrectly to be zero, it is not surprising that the other rates associated with the level of MDM2 are also inaccurate (k_3 and k_4).

One problem with using the Nelder-Mead solver is that when the initial parameter estimates are negative, the algorithm was not convergent at any ω i.e. the difference between the splines did not get below 10^{-8} (see equation C.2). Despite this the spline difference values are reasonably low (around 0.02), the residuals are low and the parameters are the right order of magnitude. Therefore the exit condition was altered to a spline difference of 0.02. The same effect was observed when Powell's optimisation method was used (appendix B.3).

If using linear algebra as the solver in Algorithm 4 predicts negative parameter estimates, the Nelder-Mead downhill simplex can be used as an alternative. The estimates will not be hugely improved as there is no additional information in the data and it is likely that the simplex routine will move the negative estimates to near zero values. A disadvantage of using the Nelder-Mead method is that it takes considerably longer to find a solution than using linear algebra and it can be less accurate. Therefore, it is advisable to only use the Nelder-Mead method when absolutely necessary.

7.7.5 Discussion

The performance of the parameter estimation routine when there is a low amount of data can be improved by increasing the number of equations, reducing the number of variables in the problem, and increasing the accuracy of the spline solution. The first two improve the "data" to variable ratio allowing the routine to produce more confident estimates, the more over-defined a set of equations is the more likely that reliable robust solutions will be obtained. The last area improves the internal estimate of the model solution. The model solution was improved by allowing the weight placed on an accurate solution to be varied and by allowing the restriction on the number of B-splines in a spline to be removed. In both cases dramatic improvements can be made in the accuracy of the

Table 7.16: The estimated parameters produced for two data sets separately and combined. There were 6 time points, 8 B-splines and 17 collocation points. ω was set to the maximum value that gives convergent behaviour; 2.743 for the 0.5Gy data, 2.25 for the 5Gy data and 2.21 for the combined data.

	D_{ATM}	D_{MDM2}	D_{p53}	p_{MDM2}	p_{p53}	k_1	k_2	k_3	k_4
True	0.05	0.18	0.041	0.01	0.52	1.42	0.39	2.5	0.75
0.5Gy	0.0500	−0.0938	0.874	−0.0638	0.497	−0.225	0.369	0.339	0.104
5Gy	0.0500	0.209	−0.457	0.131	0.328	1.87	0.244	0.0719	0.0416
Combined	0.0500	0.00902	−0.105	−0.00329	0.364	1.23	0.271	0.130	0.0329

parameter estimates.

Increasing the number of equations has been examined by increasing the number of collocation points and not restricting all data to appear at the same time points. Increasing the number of collocation points in particular led to a great improvement in the parameter estimates as the model had to be satisfied at more than a few points leading to a better model solution. Allowing data sets despite missing points is a useful feature so that data is not wasted. Another possible way to increase the amount of data and hence the number of equations is to allow data from more than one situation. For example, using data from cells that have been exposed to two different levels of DNA damage. The parameters are the same in both situations but the initial conditions are different, for example the initial DNA damage. Another set of splines needs to be introduced so that one set represents each situation. This increases the number of variables for the system but there is also an increase in the amount of data so there is a net gain in the data to equation ratio. The algorithm was easily adapted to take into account these changes. As an example of this another data set was constructed with the same parameters but with active ATM levels starting at 5 rather than 0.5. The estimated parameters are still not excellent but there is a slight improvement when both data sets are used in combination, the combined run had $A = 6.26$ (see equation 7.5) whereas separately the 0.5Gy and 5Gy runs had $A = 8.57$ and 6.92 accordingly (Table 7.16). The Y-score for combined run was better for the 5Gy data $Y = 0.00142$ instead of 0.0189 and worse for the 0.5Gy data with $Y = 0.00969$ instead of 0.00257. The limiting factor in producing better behaviour is the number of B-splines which are too low to represent the dynamics, particular the sharp dynamics of the 5Gy data. Combining two data sets could also help stabilise the data at larger numbers of splines.

Finally, if possible, the number of parameters in the model should be reduced so that the data to variable ratio improves. This could be by simplifying the model or preferably by direct measurement of parameters. For this project D_{ATM} can be accurately estimated from H2AX data (see section 3.3). Therefore this parameter can be fixed when performing the parameter estimation on real data.

7.8 Applying the collocation parameter estimation method to real p53 data

7.8.1 Introduction

Now that a new parameter estimation technique has been developed, it can be applied to the protein data gathered in the DNA damage experiments (see chapter 3).

Input to the system

The input to the system is active ATM which is assumed to be proportional to the amount of DNA damage. It is known that the rate constant for DNA repair is 0.5 hr^{-1} and that the number of initial breaks is proportional to the amount of γ radiation that the cells are exposed to (see section 3.3). Therefore, the input to the system is fully defined. For convenience the amount of active ATM can be defined in units of Gys, with the initial amount of active ATM being 5Gy or 0.5Gy.

Data quality

Ideally, so that the parameter estimation method can work accurately, there should be an absolute measure of the amount of protein for each component of the model at a large number of time points. Additionally, each data point should have a low associated error. Unfortunately, the data was very limited due to a lack of resources (see chapter 3).

The protein data gathered from the Typhoon method was used for the total p53 time course as it is the most accurate quantification method. For the other components it was arbitrarily chosen to use the first measurement (Figure 7.26). Multiple measurements were not included because measurements on different Western blots showed considerable error in the quantification (section 3.4 and 3.6). Fortunately, the 0.5Gy and 5Gy blots were always measured at the same time and on the same gel so error was minimised. Despite this there is still error in the data, for example the initial level of MDM2 is ≈ 0.6 for the 0.5Gy dose and ≈ 1 for the 5Gy dose. There is no biological reason why these should not be the same.

The data is not quantified to an absolute value but relative to the level of a standard 6 hours after 5Gy of damage. This means that the different components cannot be directly compared in this specific case because the components and therefore the parameter values are in different units. This restricts the conclusions that can be made about the biological mechanisms.

Protein measurements were only available for total p53 and active p53 whereas the models are in terms of inactive p53 and active p53. It is necessary that total p53 and active p53 are in the same units so that the amount of inactive p53 can be estimated. It is known that at the peak of the DNA damage response that the majority of p53 is in the active form, therefore it was assumed that 70% of the total p53 5Gy 6 hour standard is

active p53. The estimated inactive p53 time course is very similar to the total p53 after 0.5Gy of damage but at 5Gy there is a significant differences due to the majority of p53 becoming activated (Figure 7.26). Similarly protein measurements were only available for total MDM2, not the functionally active MDM2 used by the models. It was assumed that total MDM2 was a good approximation of the level of active MDM2 during the response.

Method

The parameter estimation routine was run with different numbers of collocation points (up to 50) and B-splines (6 to 24). In each situation the maximum weight was determined (to a resolution of 0.01) to ensure the routine produces the most accurate results. Both the 5Gy and 0.5Gy data sets were used simultaneously to ensure that the parameter values that best explain both time courses are found.

7.8.2 The simple four component model

The new parameter estimation technique was initially applied to model 1 described in section 4.3.3 (equation 4.1). The important features are that there is no inactivation of p53, no MDM2 self-ubiquitination and MDM2 ubiquitinates both forms of p53. The scores that will be used to measure the quality of the parameter estimate are,

1. The Y-score (equation 7.8) that measures how well the splines satisfy the model.
2. The least squares score, the sum of the squared distance between the spline (and hence the model solution) and the data point,

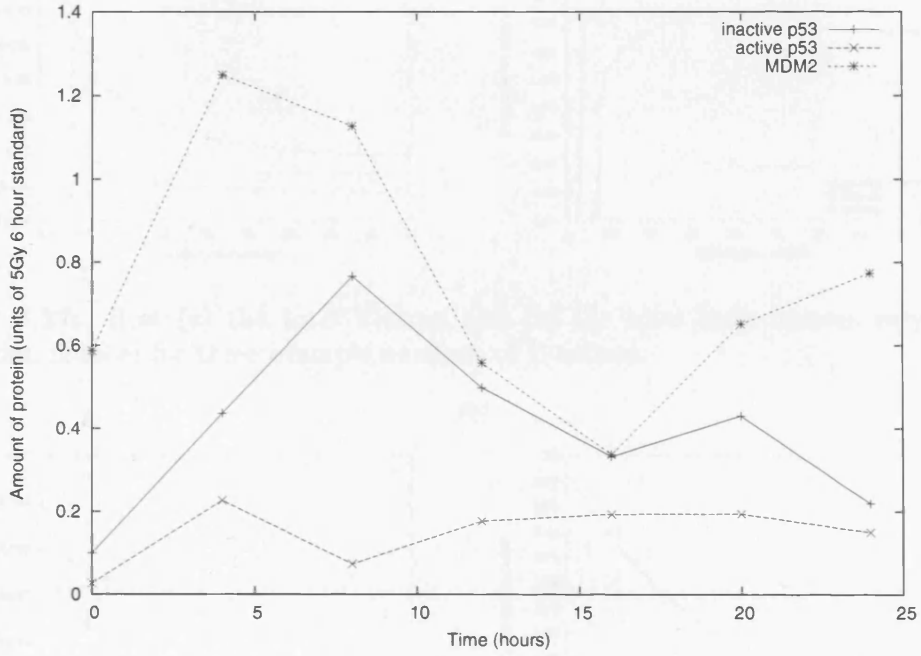
$$\sum_{i}^{n_t} \|\hat{\mathbf{x}}_i - \mathbf{x}(t_i)\|^2,$$

where $\hat{\mathbf{x}}_i$ is a vector of data at $t = t_i$ and $\mathbf{x}(t)$ is the model solution value at time t.

In both cases the lower the score is the better. As there are two data sets there is a Y-score and a least squares score for each set but normally these will be combined to give total scores for the whole parameter estimation.

As the number of collocation points increases the Y-score decreases whatever the number of B-splines, but the rate of improvement slows until some limiting value is reached (Figure 7.27(a)). An increase in the number of collocation points spreads where the spline is required to agree with the model and so overall the spline is a more accurate representation of the model. As the collocation number increases, the least squares score generally worsens, with the score reaching an almost constant value at larger numbers of collocation points (Figure 7.27(b)). The increase is caused by the improvement in the spline representing the model solution; there is less freedom in the spline to move close

(a)



(b)

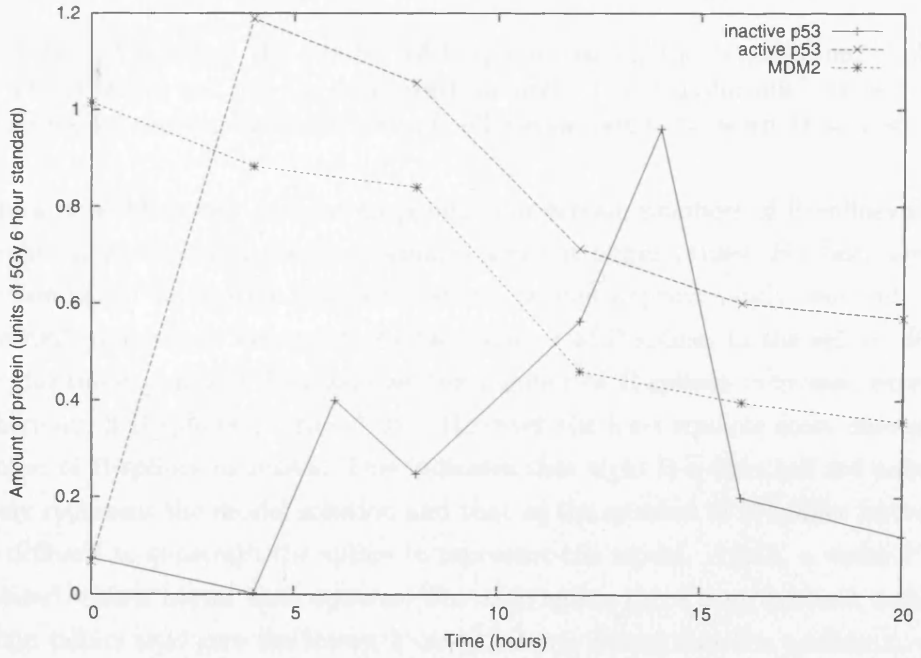


Figure 7.26: Time courses of the Western blot data after (a) 0.5Gy and (b) 5Gy of DNA damage, used for parameter estimation. Inactive p53 is in units of active p53 standard.

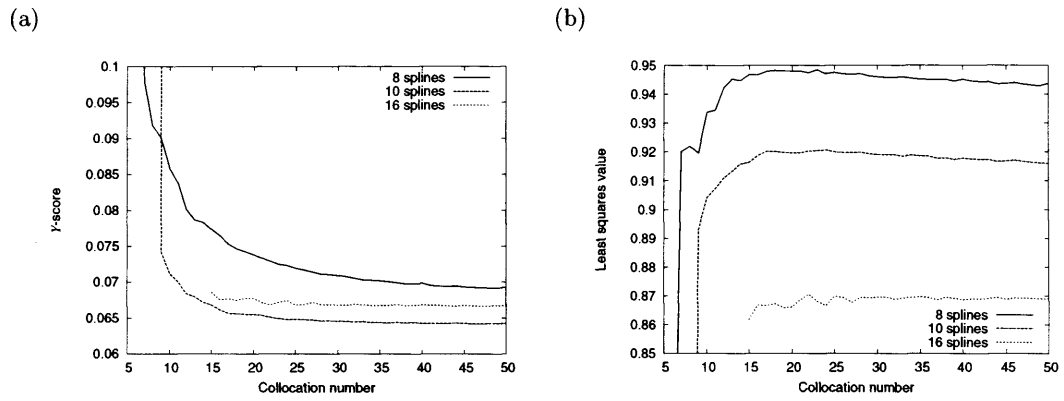


Figure 7.27: How (a) the total Y-score and (b) the total least squares vary with collocation number for three example numbers of B-splines.

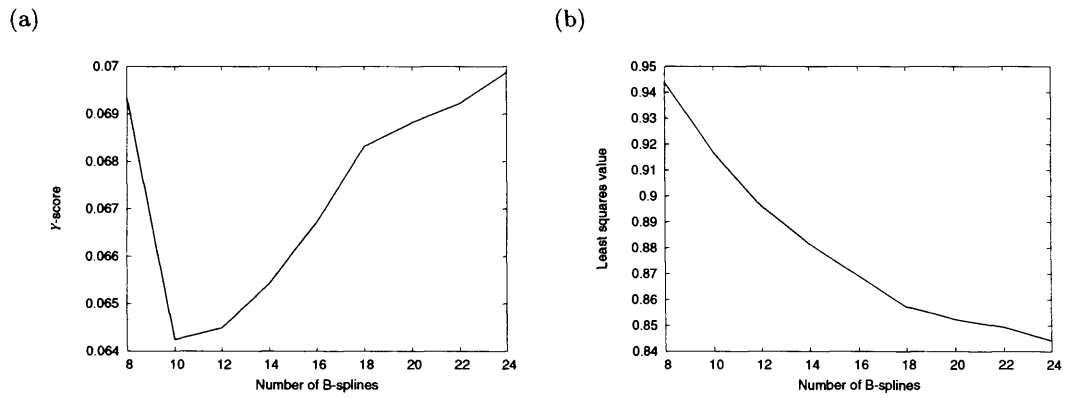


Figure 7.28: The effect the number of B-splines has on the "equilibrium" values of (a) the total Y-score and (b) the total least squares. The "equilibrium" values are the average scores for the runs with number of collocation points between 41 and 50.

to the data points between collocation points. For certain numbers of B-splines there is some slight improvement in the least squares score at larger values. For both scores at a larger number of collocation points the scores remain approximately constant. These "equilibrium" values vary depending on the number of B-splines in the spline. For the Y-score the value generally increases as the number of B-splines increases, except for when there are 8 B-splines (Figure 7.28). However the least squares score decreases as the number of B-splines increases. This indicates that eight B-splines are not enough to accurately represent the model solution and that as the number of B-splines increases it is more difficult to constrain the spline to represent the model. Again, a worse Y-score is associated with a better least squares. For each spline there is an optimal number of collocation points that give the lowest Y-score. It was found that the maximum weight for these optimal situations all remain in the same region (between 10 and 11.2) and there does not seem to be any relationship between the maximum weight and the number of B-splines used (Figure 7.29).

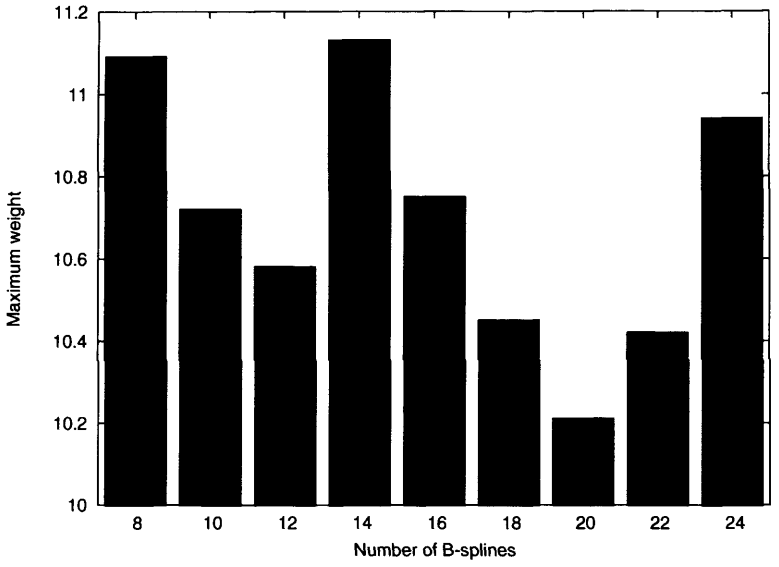


Figure 7.29: The effect the number of B-splines has on the maximum weight.

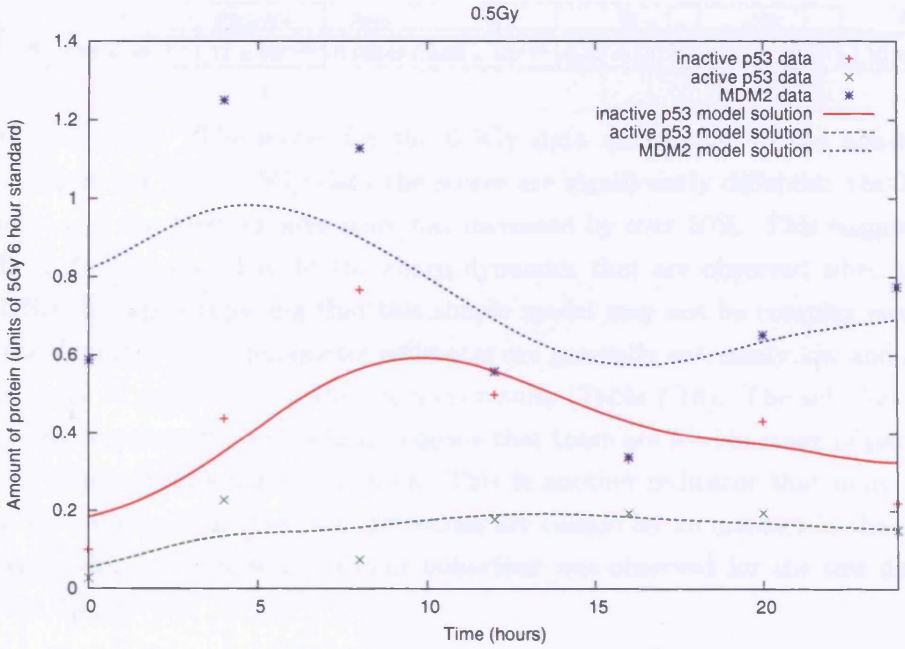
Table 7.17: The parameter estimates for model 1 (equation 4.1) found when the linear collocation parameter estimation method is applied to both 5Gy and 0.5Gy data sets. 10 B-splines per spline and 46 collocation points were used.

d_{MDM2}	d_{p53}	p_{MDM2}	p_{p53}	k_1	k_2	k_3	k_4
0.0286	0.279	0.0271	0.0153	−0.390	0.163	−0.0634	−0.0156

10 B-splines per spline and 46 collocation points give the lowest Y-score. This Y-score is reasonably low with a score of 0.0360 for the 0.5Gy data set and 0.0282 for the 5Gy data set. The associated least squares score are 0.362 and 0.555 respectively. The general shape of the resulting solution spline conforms to the data points but this model does not pick up on the fine detail suggested by the data points (Figure 7.30). In particular with the 5Gy data set the model cannot replicate the sharp increase in active p53 between 0 and 4 hours and the corresponding suppression of inactive p53. Whether the fine structure occurs in the cell or the model is inadequate depends on the amount of error in the data. To determine this would require further data.

Table 7.17 displays the resulting parameter estimates. As mentioned above due to the parameters being in units that are not comparable it is difficult to tell much about the mechanisms from the parameter estimates. k_1 (rate MDM2 ubiquitinates p53), k_3 (rate active p53 transcribes MDM2) and k_4 (rate active ATM inactivates MDM2) have been estimated to be negative. It is conceivable that the mechanism of active p53 and ATM has been reversed but this does not explain why k_1 is negative. The parameter estimation method was repeated with the simplex solver (section 7.7.4) so that only positive parameter values are valid (10 B-splines per spline and 46 collocation parameters were used). The maximum weight factor was 10.6 and the Y-score was 0.0393 for the 0.5Gy data and 0.0591 for the 5Gy data set. The corresponding least squares scores

(a)



(b)

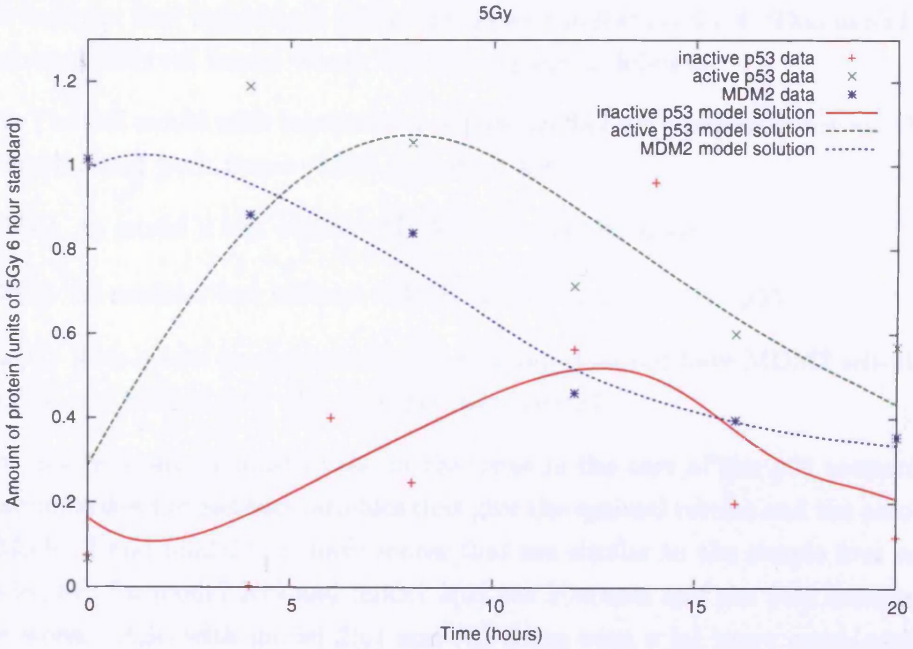


Figure 7.30: The model solution for (a) 0.5Gy and (b) 5Gy data for model 1 (equation 4.1) when the optimal settings were used.

Table 7.18: The parameter estimates for model 1 (equation 4.1) when the simplex solver was used. Both 5Gy and 0.5Gy data sets, 10 B-splines per spline and 46 collocation points were used.

d_{MDM2}	d_{p53}	p_{MDM2}	p_{p53}	k_1	k_2	k_3	k_4
3.88×10^{-34}	7.63×10^{-9}	1.12×10^{-36}	0.00549	8.02×10^{-32}	8.47×10^{-28}	3.36×10^{-30}	1.10×10^{-29}

are 0.385 and 0.825. The scores for the 0.5Gy data are similar to the non-simplex implementation but for the 5Gy data the scores are significantly different: the Y-score has doubled and the least squares score has increased by over 50%. This suggests that it is difficult for the model to fit the sharp dynamics that are observed when there is 5Gy of DNA damage suggesting that this simple model may not be complex enough to explain the dynamics. The parameter estimates are generally extremely low and are not the same order of magnitude as the previous results (Table 7.18). The solution splines still fit the data reasonably well which suggests that there are a wide range of parameter values that produce reasonable dynamics. This is another indicator that more data is required. It is also possible that the low values are caused by an artefact in the method caused by moving to a log scale; similar behaviour was observed for the test data sets (see section 7.7.4).

7.8.3 The complex four component model

To further investigate the p53 system the parameter estimation method was applied to the more complex four component model introduced in section 4.3.4. This model comes in a number of different forms, whose key features are as follows,

model 2 The full model with inactivation of p53, MDM2 self-ubiquitination and MDM2 ubiquitinating both forms of p53 (equation 4.2).

model 2(b) As model 2 but without MDM2 self-ubiquitination.

model 2(c) As model 2 but without MDM2 ubiquitinating active p53.

model 2(d) This model has inactivation of p53, but does not have MDM2 self-ubiquitination and MDM2 only ubiquitinates inactive p53.

These models examine most of the mechanisms in the core of the p53 network. Table 7.19 summarises the method variables that give the optimal results and the associated scores. Model 2 and model 2(b) have scores that are similar to the simple four component model, but for model 2(c) and model 2(d) the Y-scores and the 5Gy least squares score are worse. Also with model 2(c) and (d) there were a lot more combinations of numbers of collocation points and numbers of B-splines that would not converge for any weight factor. This suggests that the ubiquitination of active p53 might be an important mechanism whereas MDM2 self-ubiquitination and inactivation of p53 might be less important. Unfortunately, there is not enough accurate data to give strength to this

Table 7.19: Summary of the results from parameter estimation on various complex four component models.

model	Number of B-splines	Number of collocation points	Maximum weight	Y-score (0.5Gy)	Y-score (5Gy)	least squares score (0.5Gy)	least squares score (5Gy)
4	10	46	10.72	0.0322	0.0205	0.381	0.542
4b	10	46	10.72	0.0320	0.0203	0.380	0.541
4c	8	48	10.47	0.0419	0.0544	0.384	0.701
4d	8	48	10.45	0.0420	0.0546	0.383	0.699

Table 7.20: The parameter estimates for the complex four component p53 model.

model	d_{MDM2}	d_{p53}	p_{MDM2}	p_{p53}	k_1	k_2	k_3	k_4	k_5
4	0.0164	0.258	0.0154	−0.0432	−0.460	0.165	−0.0615	−0.0140	0.0669
4b	0.0283	0.255	0.0270	−0.0439	−0.458	0.163	−0.0631	−0.0150	0.0676
4c	0.0170	−0.0216	0.0157	−0.0344	−0.0869	0.212	−0.0668	−0.0180	0.0369
4d	0.0320	−0.0172	0.0301	−0.0278	−0.0747	0.220	−0.0697	−0.0199	0.0336

assertion. The estimated parameters (Table 7.20) are reasonably consistent across all the models. Exceptions are the production rate of p53 (p_{p53}) which is negative for the "complex" models but not model 1 and the degradation rate of MDM2 (d_{MDM2}), which is significantly lower when MDM2 ubiquitinates itself. The wide spread negative values suggests that either the models are not appropriate or that there is insufficient data for the parameter estimation method to be accurate.

7.9 Conclusion

In this chapter an interesting new approach to parameter estimation based on linear algebra, collocation and B-splines has been developed. It is based around using a spline as an intermediary between the model solution and the data. By using linear algebra a spline can be found that maximises its closeness to the data points and its satisfaction of the model. As a result parameter estimates are made. This method accurately predicts the parameter values, especially when the error in the data is low.

When the error is large it is no longer satisfactory for the spline to balance the distance away from the data points and its ability to represent the model solution because the spline would inaccurately represent the model. Therefore a refinement was made so that the spline was pressurised to more closely represent the model by introducing a weight factor, ω, that varies how much notice is taken of the data. As ω is increased the intermediary spline more closely represents the model and produces better parameter estimates. If ω is increased too far though, the algorithm becomes divergent as too little attention is applied to the data. This refinement produces better parameter estimates. It has however introduced an additional problem of finding the optimal ω that still allows

the algorithm to converge. In some situations it is not possible to use this parameter estimation method as a valid ω cannot be found. A by-product of this refinement is that there is no constraint on the number of collocation points. Increasing the number of collocation points generally improves the estimate, but if it is increased too far a slight deterioration in performance can occur due to numerical errors.

When there is a low amount of data it is very difficult if not impossible to produce decent parameter estimates in any estimation routine. Various refinements have been proposed that potentially improve the algorithm when there is a small amount of data. Three areas were investigated, increasing the number of equations, reducing the number of variables in the problem and increasing the accuracy of the spline solution. It was found that definite improvements were made by increasing the number of collocation points, B-splines per spline and ω. None of these refinements guarantee accurate parameter estimates but they can help to get reasonable parameter estimates that are good enough to use in my situations.

This parameter estimation method has some advantages over alternative techniques. Firstly, it is a fast algorithm, this arises because the problem has been linearised so fast linear algebra solvers can be used and by using splines the model has effectively been discretised in an effective way avoiding the need for costly and accurate integration of the model. For data with a small amount of error the algorithm is extremely accurate despite the high dimension of the problem. Other methods such as the Nelder-Mead downhill simplex method and Powell's direction set method are particularly prone to numeric error at high dimensions. The parameter estimation method is also reasonably simple with only three key parameters that can be varied compared to simulated annealing where there is a numerous algorithm parameters that can be tweaked. The key is that it works for the models proposed in this thesis where other parameter estimation methods have failed.

There are a number of important restrictions to this parameter estimation procedure. For the algorithm to provide reasonable parameters, an ω has to be found so that the algorithm is convergent and gives a low Y-score (the model is being accurately represented by the spline). This may not be possible in some cases and so the parameter estimates can not be trusted in these situations. Another restriction is that the model has to be linear in its parameters (see section 7.2.3). This is a fundamental requirement of the algorithm. It is difficult to see how this restriction could be overcome but one possibility would be to fix all the non-linear parameters apart from one, run the algorithm, and then fix the free parameter and make another one free and run the algorithm. This could be repeated until there was convergence in the parameters. There also might be limited scope in performing a power expansion on the non-linear parameters, but it is difficult to see how one would deal with a straight product e.g. $a \times b$.

There are still many areas where this algorithm could be further developed and analysed. In particular it would be useful to get a more concrete grasp on what is causing the

spline to diverge in some situations rather than others. Even though a hypothesis has been put forward, this area needs to be tested and examined in more detail. Interlinked with this is the way that the equations are re-weighed to get an improvement in the spline representing the model solution. At the moment a fairly simplistic approach is taken with the whole block of data-spline and model-spline equations being re-weighed by the same amount. Weighing individual equations could cause an improvement. In particular, the larger the value of the data point the greater the weight it has, this means that if a component has a very low level it will not be as likely to have a good model fit. A way round this would be to re-scale each data equation so that they all equal a constant value, which would have the effect of changing the algorithm to a relative error approach. It would also be interesting to examine the performance of the algorithm as error is increased in a small amount of data, the algorithm would be expected to fail very quickly. This leads to another question of how much data is required to get reasonable parameter estimates.

When the linear collocation parameter estimation technique was applied to experimental data and a variety of p53 models there is the suggestion that the ubiquitination of active p53 is an important mechanism in the system whereas MDM2 self-ubiquitination and p53 inactivation are not. Unfortunately it is impossible to give strength to these assertions due to the limited data set available. From pseudo data it is known that the parameter estimation method becomes increasingly unreliable as the amount of data is decreased. The amount of data that was available was close to minimum required for this technique and so it is not surprising that it performed poorly. Other parameter estimation techniques would have equal if not more difficulty. Another difficulty with the data was that it was in relative units which prevented the comparison of the parameters and hence it was impossible to discuss the relative importance of each mechanism. Gaining absolute concentrations by quantifying the amount of protein in the standard appears to be essential. Due to a lack of resources only data from the core components were gathered so it was impossible to examine models with a more diverse range of components. Finally, the technique used to quantify the majority of the data was prone to large errors and unfortunately there was not the resources to perform repeat experiments on an improved method (chapter 3). Having an accurate measure of error in the data would help considerably in the evaluation of models, especially if this information could be used by the parameter estimation technique. Although this investigation has not proved conclusive, the problems are surmountable and in the future, with technological and computational advances, the quantification of diverse protein data sets will become accurate and reliable. Despite these results, the validity of the parameter estimation technique is not in doubt due to its performance on theoretical data sets.

Chapter 8

Construction of transcription activity profiles and their applications

8.1 Introduction

So far in this thesis the focus has been on the regulation of p53 and how this changes after DNA damage. This is only one part of the p53 gene regulatory network and more generally the DNA damage response. In this chapter tools will be developed to examine a more global view of the response, with a particular focus on the targets of p53.

8.1.1 The biological problem

Microarray experiments measure the mRNA levels of thousands of genes simultaneously and can be used to gain insight into the function of gene regulatory networks. To understand a gene regulatory network it is necessary to know its composition and obtain a quantitative description of how the components interact with each other at both the protein and mRNA level. Microarray data only provides mRNA level information and it is common to assume that genes with closely related expression patterns are controlled by the same regulatory mechanisms ("guilt by association") (Schulze and Downward, 2001). This approach ignores key factors that will affect the mRNA level, such as the effectiveness of a transcription factor and the rate at which the mRNA decays. In addition, there is no quantitative measure of how likely a particular gene is to be regulated by a particular transcription factor. An alternative approach is to use a mathematical model to provide a link between mRNA levels and the activity driving the transcription. Dynamic modelling of microarray time course data takes into account parameters ignored by standard analysis methods and can extract hidden information about the transcription activity profile and apply this to make quantitative predictions about network behaviour. In theory, this type of modelling can identify genes that share the same transcription factor and derive the activity time profile of that transcription factor. Ultimately, modelling has the potential to fully reconstruct the gene regulatory network.

In this chapter a method will be presented that uses time series microarray data and a simple model of gene transcription to construct a quantity that can be used to predict the transcriptional response of a cell population to DNA damage. In particular it will be used to examine transcription factor activity profiles and to identify p53 and other transcription factor targets.

8.1.2 Barenco *et al.* hidden variable modelling

In a submitted publication from this group, Barenco *et al.* (2005) produced interesting work in this area. The general approach is to predict transcription factor targets from microarray data by deriving and exploiting "hidden variables". The hidden variables are factors that influence the data but are not directly measured, in this case the main hidden variables are the activity profile of a transcription factor and the gene target's sensitivity to that activity. These hidden variables are predicted by using a simple model of gene transcription based on prior biological knowledge (see section 8.1.3). The example system

studied is the transcriptional response to ionisation radiation with the aim to recover p53 targets. The experiments produced microarray time series (see section 3.2.2). The first step is to use five known p53 target genes to derive the p53 transcriptional activity profile. This is done by finding the hidden variables of the model that allow the best fit to the data using Markov Chain Monte Carlo (MCMC) with a Metropolis-Gibbs sampler. The parameters include the discretised transcription factor activity profile which is assumed to be shared by all the training target genes. To speed up the solution of the model, the differential operator was discretised (see section 8.2.1). The resulting transcription profile was found to be in good agreement with experimental data.

In the second step, the transcription factor activity profile is fixed, changing the model to a model of p53 target gene transcription. This model is then used to screen all up-regulated genes to identify likely p53 targets. This is done by estimating the remaining parameters of the model that produce the best fit to the data. Two relevant scores arise from this process; the model score (how well the model describes the data, the lower the better), and the sensitivity (how sensitive the gene's mRNA concentration is to changes in p53 concentration, the higher the better). Based on these scores the up-regulated genes can be divided into three classes: those predicted to be p53 targets - low model score and high sensitivity, those predicted to be co-regulated by p53 or are independent of p53 - high model score and high sensitivity, and those that are independently regulated - high model score and low sensitivity. To validate the findings microarrays were run on MOLT4 cells transfected with siRNAp53 after exposure to radiation (see section 8.4.4). It was found that the top 50 genes predicted to be p53 targets by the model were highly sensitive to siRNAp53 confirming the use and validity of this approach.

The work in this chapter uses a similar approach to that used by Barenco *et al.* (2005). The parts in this work that are directly taken from Barenco *et al.* (2005) are the model of gene transcription (see section 8.1.3) and the procedure used to discretise the differential operator (see section 8.2.1). In addition to this the same gene transcription time series data and p53 verification data is used. This work though is a novel approach to the same problem; instead of performing parameter estimation to extract the transcription activity profile, here one parameter is determined by biological experiment and the data is manipulated in such a way that the transcription activity profile can be considered directly. This approach is more versatile than Barenco *et al.* (2005), with the discovery of target genes based on a training set of genes as only one possible application. It can also link genes that share the same transcription activator profile without any prior knowledge of target genes and also extract global information about the response, for example the number of strong transcription activation profiles that are driving the system.

8.1.3 A simple model of gene transcription

A *transcription activity* is any effect or combination of effects that regulates the production of a gene's mRNA[1]. The transcription activity profile, $f(t)$, determines the rate, at time t, at which the concentration of mRNA, x_g, of gene g is produced. It is assumed that the rate of production is linearly proportional to the level of the transcription activity with the constant of proportionality, S_g. S_g describes the sensitivity that the activity has on the amount of mRNA. In this study it is assumed that S_g is positive i.e. only transcription activities that have a positive regulatory effect are considered. There are two other factors that affect the amount of mRNA; the basal transcription rate B_g, which is the rate of production of mRNA in the absence of the transcription activity, and the degradation rate of mRNA which is assumed to be proportional to the level of mRNA, the amount lost per unit time is $D_g x_g(t)$, where D_g is the degradation rate constant of gene g. A linear differential equation that summarises this behaviour is,

$$\frac{dx_g(t)}{dt} = B_g + S_g f(t) - D_g x_g(t). \tag{8.1}$$

8.1.4 The G time series

It would be interesting to know what genes share the same transcription activator profile i.e. those genes whose transcription is regulated by the same activity. To this end it is assumed that the transcription model (equation 8.1) holds[2], and a quantity $G_g(t)$ is proposed for each gene g,

$$G_g(t) = B_g + S_g f(t), \tag{8.2}$$

which is equivalent to (using equation 8.1),

$$G_g(t) = \frac{dx_g(t)}{dt} + D_g x_g(t). \tag{8.3}$$

The $G_g(t)$ time course can be calculated using equation 8.3, as $x_g(t)$ is known, $\frac{dx_g(t)}{dt}$ can be estimated from $x_g(t)$ and the degradation rates can be measured.

As equation 8.2 shows, $G_g(t)$ is a simple linear transformation of the transcription activator profile, with B_g shifting $G_g(t)$ a fixed amount away from $f(t)$ and S_g stretching $G_g(t)$ relative to $f(t)$. This means that any two genes that share the same transcription activator profile will have $G_g(t)$s that are perfectly correlated (see appendix C.5 for proof). $G_g(t)$ represents the transcription activity profile of gene g, as long as correlation is used as the distance measure when comparisons are made.

[1]The simplest transcription activity would be the amount of a functionally active transcription factor. It could also be the activity of a complex of proteins that combine to provide transcription functionality. It is even possible that a single transcription factor might have two different transcription activity profiles.

[2]Even though the transcription model (equation 8.1) is limited because it assumes that all effects are linear and cooperative effects between two or more transcription activator profiles are ignored, it still provides a useful link between the mRNA and transcriptional activities.

In reality the model will not describe the system perfectly and the data will have errors, therefore if there is a reasonably good correlation between two genes' $G_g(t)$ time profile then it is likely that the two genes share the same transcription activator profile. If two transcriptional activity profiles are the same or similar then this approach will not be able to distinguish between them. Also, the effect of the transcription activator profile would have to be strong for its effects to be detectable.

The traditional approach to examining microarray data is restricted to the gene expression profiles. The conjecture is made that if two gene expression profiles share the same shape then they share some common function or regulation. This is only reasonable if the mRNA degradation rates are similar, because the shape of the expression profile is determined by the degradation rate (Figure 8.1). Therefore, the distribution of degradation rates determines what clusters are formed. Two clusters could be formed by the same activity profile and some targets of a particular activity profile will be overlooked. The G time series removes the degradation rate as an influence and gives the shape of the transcription activity profile. This should improve the quality of results.

8.2 Constructing the G time series

The $G_g(t)$ time course can be calculated using equation 8.3. This requires the degradation rates of all genes to be measured and a way to estimate the rate of change of expression level from time series microarray data. These will be examined below. After all the terms of $G_g(t)$ are found, simple matrix algebra can be used to calculate the $G_g(t)$ time course.

8.2.1 Estimating the rate of change of gene expression level

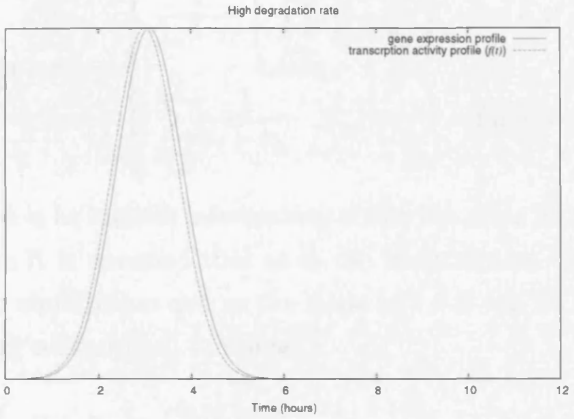
The estimation of $\frac{dx_g(t)}{dt}$ from the microarray data will be calculated using the same method as Barenco *et al.* (2005). This method is motivated by collocation based approaches to boundary value problems for non-linear differential equations (Golub and Ortega, 1992) and converts the derivative term into an algebraic one. Suppose that the transcription level of gene g, $\hat{x}_g$ is observed at $n+1$ time points $t_0, t_1, \ldots, t_n$. The derivative of the expression profile at time t_i is estimated as follows,

$$\frac{dx_g(t_i)}{dt} \approx \sum_{k=0}^{n} A_{ik}\hat{x}_g(t_k),$$

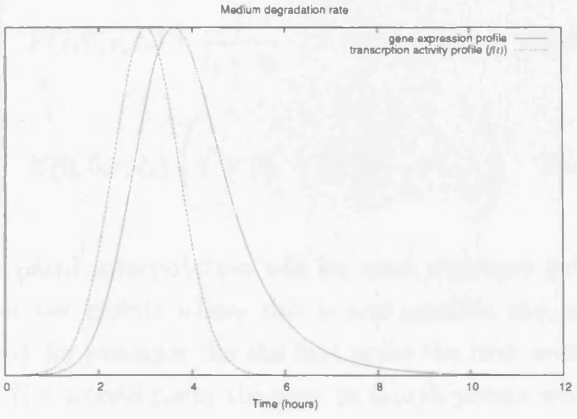
where A is an $(n+1)\times(n+1)$ matrix. The basic method to find the values of A is to fit a $q-1$ degree polynomial to q points around the point of interest by Lagrange interpolation and then differentiating the polynomial. Let p be the initial point considered and r the end point ($r = p+q-1$), then the coefficients of the A matrix will be,

$$A_{ik} = \begin{cases} E(k,p,r,t_i) & \text{for } p \le k \le r \\ 0 & \text{otherwise} \end{cases}$$

(a)



(b)



(c)

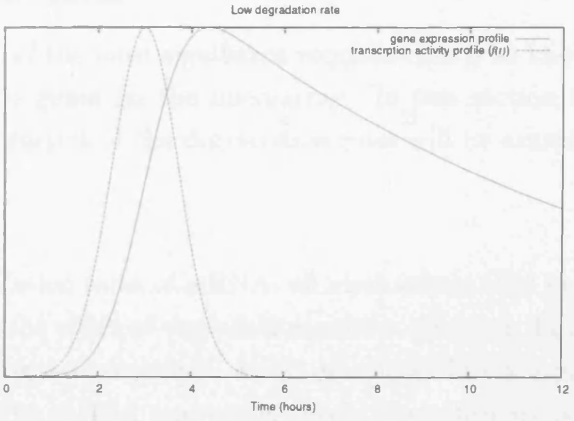


Figure 8.1: A plot to show how the shape of the expression profile depends on the degradation rate and how the higher the degradation rate, the closer the profile is to the transcription activity profile. This is based on equation 8.1 with (a) $D_g = 10$,(b) $D_g = 1$ and (c) $D_g = 0.1$. $f(t) = \exp(-(x-3)^2)$, $S_g = 2$ and $B_g = D_g$ (so that the equilibrium level is at one). Expression levels have been scaled to have the same height as $f(t)$.

where,

$$E(k,p,r,t_i) = \begin{cases} \dfrac{1}{t_k - t_i}\left(\prod\limits_{\substack{j=p \\ j\neq i,k}}^{r} \dfrac{t_i - t_g}{t_k - t_g}\right) & \text{for } k \neq i, \\ \sum\limits_{\substack{j=p \\ j\neq i}}^{r} \dfrac{1}{t_i - t_g} & \text{for } k = i. \end{cases}$$

A refinement of this is to include information about the slope at t_0 when the t_0 point is used in the fit. Here it is assumed that at t_0 the transcription factors and the genes that it regulates are in equilibrium and so the slope at $t = 0$ can be assumed to be zero. This gives the following refined $E_0()$ formulae,

$$E_0(k,p,r,t_i) = \begin{cases} E(k,0,r,t_i)\dfrac{t_0 - t_i}{t_0 - t_k} & \text{for } k \neq 0, i, \\ E(i,0,r,t_i) + \dfrac{1}{t_i - t_0} & \text{for } k = i, k \neq 0, \\ E(0,0,r,t_i)\left(1 + (t_0 - t_i)\sum\limits_{j=1}^{r} \dfrac{1}{t_0 - t_g}\right) & \text{for } k = 0. \end{cases}$$

Here a symmetric five point interpolation will be used wherever possible. For the first two points and the last two points where this is not possible the points that go out of range are simply ignored, for example, for the first point the first, second and third points will be used whilst for the second point the first to fourth points will be used for the fit.

8.2.2 Degradation Rates

To calculate $G_g(t)$ one of the most significant requirements is to know the mRNA degradation rates for all the genes on the microarray. In this section the method used to produce reasonable estimates of the degradation rates will be examined.

The experiment

To measure the degradation rates of mRNA, all mechanisms that produced mRNA need to be stopped, so only the effect of degradation can be observed. In human MOLT4 cells this was done by adding actinomycinD, which effectively blocks gene transcription (Alberts *et al.*, 2002). The mRNA expression levels were then measured by microarray experiments as the levels decay away.

The experiment was performed on MOLT 4 cells that had been irradiated so that expression levels of genes important to the DNA damage response are detectable above the background noise, providing a more accurate estimate of the degradation rates. The cells in log phase (10^6 ml^{-1}) were γ-irradiated with 5 Gy at room temperature at a

dose rate of 2.45 Gy per minute (using a 137Cs g-irradiator). The cells were then left for four hours so that the peak response was reached. The cells were then split into two batches with one half having 10 μg ml^{-1} actinomycinD added (this is the zero hour point). RNA and protein were extracted at 0, 0.5, 1, 2, 3, 4 and 6 hours. RNA and cRNA were prepared, and their quantity and quality was determined by Nanodrop spectophotometer and Bioanalyser 2100 (Agilent). A proportion of the mRNA at each time point was used to perform microarray measurements using the Affymetrix U133A arrays (see appendix A.3). The gene expression levels were calculated using the MAS5.0 algorithm (Affymetrix, 2002a,b). MAS5.0 has limitations so adjustments need to be made to get reasonable degradation rates. This is done in two steps: re-normalising the arrays so that the smallest degradation rate is zero and anchoring the data so there is maximum agreement with degradation rates found from QPCR. Both of these adjustments require estimates of the degradation rates and how these are found are considered below.

Linear regression

To find the most appropriate degradation rate for each gene (D_g) one must minimise the difference between the model of degradation and the data. Here the weighted least squares error is used as a measure of this difference,

$$WLSQ(D_g, A_{0g}) = \sum_{i=1}^{n} \left(\frac{\hat{x}_g(t_i) - A_{0g}e^{(-D_g t_i)}}{\sigma_i} \right)^2$$

where $\hat{x}_g(t_i)$ is the gene expression level at time t_i, σ_i is the uncertainty associated with measurement $\hat{x}_g(t_i)$, A_{0g} is the initial mRNA expression value and n is the number of data points. This could be solved by an optimisation routine such as the Nelder-Mead simplex (Nelder and Mead, 1965) but it is computationally expensive. By converting to the log domain a linear fit can be used, which is fast and provides a statistical goodness of fit measure. The equation can be rearranged into a line equation where the gradient is the negative of the degradation rate constant, $\ln(A_0)$ is the y-intercept, y is the logarithm of the gene expression, and x is time;

$$\begin{aligned} x_g(t) &= A_{0g}e^{(-D_g t)}, \\ \Rightarrow \ln(x_g(t)) &= \ln(A_{0g}) - D_g t, \\ \Rightarrow y &= a + b \times x. \end{aligned}$$

The line fitting routine was implemented in C++ using equations from Press *et al.* (2002, section 15.2): if there are N data points (x_i, y_i) with associated uncertainty σ_i,

$$b = \frac{1}{S_{tt}} \sum_{i=0}^{N-1} \frac{t_i y_i}{\sigma_i}, \quad a = \frac{S_y - S_x b}{S}, \quad \sigma_b^2 = \frac{1}{S_{tt}},$$

where,

$$S \equiv \sum_{i=0}^{N-1} \frac{1}{\sigma_i^2}, \quad S_x \equiv \sum_{i=0}^{N-1} \frac{x_i}{\sigma_i^2}, \quad S_y \equiv \sum_{i=0}^{N-1} \frac{y_i}{\sigma_i^2},$$

$$\begin{aligned} t_i &= \frac{1}{\sigma_i}\left(x_i - \frac{S_x}{S}\right), \\ S_{tt} &= \sum_{i=0}^{N-1} t_i^2. \end{aligned}$$

It is assumed that the measurement errors are normally distributed and independent. The $WLSQ$ is therefore distributed according to a χ^2 distribution with $n-2$ degrees of freedom under the null model of linearity (n is the number of data points). The null hypothesis, that the data is distributed according to a linear model, can therefore be tested by computing $\mathrm{P}(\chi^2 > WLSQ)$ (this is the p-value) as follows,

$$p = \frac{\Gamma\left(\frac{n-2}{2}, \frac{WLSQ}{2}\right)}{\Gamma\left(\frac{n-2}{2}\right)}$$

where $\Gamma()$ is the gamma function (this is the upper regularised gamma function). Press *et al.* (2002) suggest that if p is less than 0.001 then linear regression is probably not appropriate.

The tailing-off effect

Due to a variety of effects it is often the case that the degradation data does not appear to decay at a constant rate. In these cases it is beneficial to remove the offending data from the fit, so a more accurate estimate of the degradation rate can be achieved. One particular problem is that the amount of expressed mRNA sometimes appears to stop decreasing (Figure 8.2). This problem will be called the *tailing-off effect*. Two possible causes are:

1. The gene expression has decayed to such a low value that any further decrease is lost in noise; microarray experiments give a large amount of error at low values of gene expression.

2. The agents of degradation (nucleases) cease to function after a certain time. There will be a decreasing population of functional degradation agents because these agents will degrade as normal but not be replaced due to the lack of transcription.

Data points that occur within the tailing-off region should not be included in the linear regression as they will have an adverse effect on the estimation of the degradation rate.

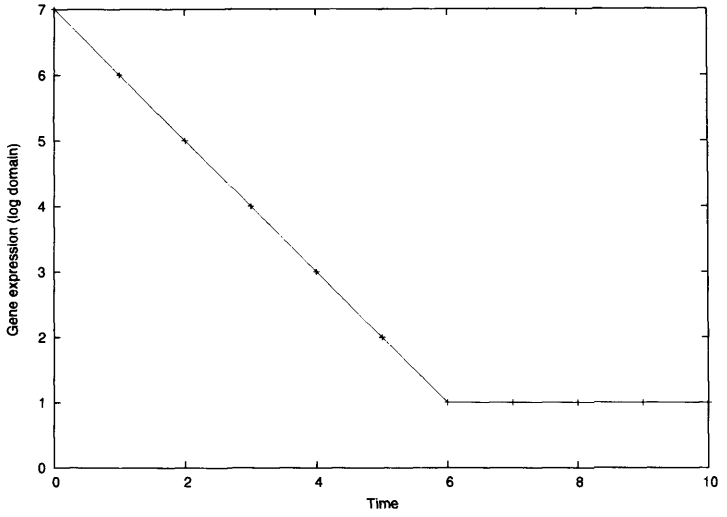


Figure 8.2: An idealised example of the tailing-off effect. The expression level decreases linearly obeying the degradation model up to 6 hours. Afterwards it stops decreasing.

Θ_i^j is defined as gradient determined by linear regression when applied to all data between time t_i and t_j, ϵ_i^j is the associated error of the gradient estimate and p_i^j is the associated p-value. If,

$$\Theta_i^j - \epsilon_i^j \leq 0 \leq \Theta_i^j + \epsilon_i^j,$$

then there is a "straight detection" between data points $\hat{x}(t_i)$ and $\hat{x}(t_j)$. To determine whether degradation data of a particular gene has the tailing-off effect, Θ_i^n is found for $3 \leq i \leq n-2$, where n is the total number of data points. If there are two or more "straight detections" or if there is a "straight detection" for $i = n-2$ than the gene has the tailing-off effect. This detection process is repeated for Θ_i^{n-1} with $3 \leq i \leq n-3$ as sometimes the final point had a significant effect on the gradient causing problems.

If the data for a particular gene is determined to have the tailing-off effect then inappropriate data is removed using the following fix. Θ_1^j is calculated for $4 \leq j \leq n$.

$$p_1^{j^*} = \max\left\{p_1^j \,|\, 4 \leq j \leq n\right\},$$

where $p_1^{j^*}$ is the largest p-value found. $\Theta_1^{j^*}$ is then used as the estimate of the degradation rate. When data within the tailing-off region is included in the linear regression, the data will not fit the linearity model very well and so the p-value will be poor (in this case low). Therefore, taking the linear regression with the highest p-value will give a data set that excludes the problem points. The nature of the p-value weighs in favour of there being more data points so this fix should only exclude those points that are within the tailing-off region. It was arbitrarily chosen that the minimum number of data points in a linear

regression was four.

This fix was also applied in other situations where the data was not ideal and steps had to be made to pick the correct amount of data. These situations were if the p-value was poor ($p_1^n < 10^{-10}$) or if the last point causes a significant change (greater than 8 times the average) in the gradient estimate.

Adjusting the linear regression when transcription has not stopped at $t = 0$

Sometimes the gene expression levels appeared to rise after the first time point, which is probably due to transcription not being completely suppressed at $t = 0$. The data set was determined to have this problem if the second data point was significantly higher than the first i.e. $\hat{x}_g(t_2) - \sigma_{g2} > \hat{x}_g(t_1) + \sigma_{g1}$, where σ_{gt} is the standard deviation of the error of gene g at time t. If this was the case then linear regression was performed with and without the first data point and the gradient of the regression that gave the best p-value was taken as the degradation rate constant. If the data has the tailing-off effect and the first point was significantly lower than the second then the tailing-off effect fix was applied with and without the first point and the degradation estimate that had the best associated p-value was used.

Re-normalisation

The MAS5.0 algorithm normalises each microarray so that they have the same total level of gene expression. For the degradation time series this normalisation is undesirable as the total level of expression is expected to drop as the mRNA decays. Therefore each microarray needs to be re-scaled to reflect this. All expression levels are re-scaled according to,

$$\hat{x}_g^{\text{new}}(t_i) = \alpha_i \hat{x}_g(t_i),$$

where α_i is the adjustment factor associated with the microarray that measures gene expression at time t_i. It is known that once gene transcription has been stopped, all gene expression levels should decay and so no genes should have an increasing gene expression. Therefore α_i was found such that the slowest decaying genes have a degradation rate of zero. The first microarray ($t = 0$) remains fixed and so α_0 has a value of 1.

Using the line fitting procedure and fixes described above the decay rates are found for all genes. The genes are ranked according to their estimated decay rate constants. The 1% of genes with the lowest decay rates (including negative decay rate constants) are referred to as the slow set. For each gene, j, in the slow set α_{ij} is found so that the time course is flat,

$$\alpha_{ij} = \frac{x_j(0)}{x_j(t_i)}.$$

Table 8.1: The degradation rate constants found from the QPCR data.

Gene	Degradation rate constant
GADD45	1.68
MDM2	0.253
p21	0.904
CD71	0.602
HPRT1	0.522
PGK1	0.531

α_i is then found by averaging the changes,

$$\alpha_i = \frac{1}{n_s} \sum_j^{n_s} \alpha_{ij},$$

where n_s is the number of genes in the slow set. The data is then re-scaled by α_i. α_is are then calculated for the new data set. This process is repeated on each new data set until no further scaling is required. It takes only a few iterations before convergence. Sometimes it converges to an oscillation between two groups of α_is, and so one of these groups is chosen at random.

Anchoring microarray data using QPCR data

The re-normalisation attempts to undo the normalisation done by MAS5.0 by adjusting each microarray separately so that the minimum degradation rate is zero. The minimum degradation rate will actually be greater than this, so after re-normalisation the degradation rates will be underestimated. One way to correct this is to re-adjust the data so that the degradation rates estimated from microarray data are in agreement with degradation rates estimated from a different source. In this case, quantitative PCR measurements (see appendix A.2) were used to anchor the microarray data to an independent estimate of a few degradation rates.

A proportion of the cells harvested in the degradation experiments were used to make QPCR measurements of the p53 target genes GADD45, MDM2, p21, CD71, HPRT1 and PGK1. All QPCR measurements were performed in triplicate. From the QPCR data and the fitting procedure described above an estimate of the degradation rate was obtained for each of the genes processed. The linear regression was performed on the average of three technical replicates and the error was set to the standard deviation of the three replicates (see Table 8.1).

A constant, β, is found such that the two sets of degradation rates are in the best possible agreement by minimising,

$$\sum_{k=1}^{m} (D_k + \beta - D_k^*)^2, \tag{8.4}$$

where D_k^* is the estimated degradation rate from QPCR of transcript k, D_k is the degradation rate estimated from microarray data, and m is the total number of Affymetrix probe sets associated with all of the transcripts measured in the QPCR experiment[3]. This minimisation is solved by performing a linear regression on the degradation rates with the gradient fixed at one (x-axis: microarray degradation rates, y-axis: QPCR degradation rates). β is set at the y-intercept. The complete microarray data set is then adjusted according to the following formula,

$$x_g^{\text{new}}(t_i) = x_g(t_i)e^{-\beta(t_i - t_0)}, \tag{8.5}$$

where $x_g^{\text{new}}(t)$ is the adjusted expression value of gene g at time point t_i and t_0 is the time at the initial time point (which is fixed). After the microarray data has been adjusted, the microarray degradation rates are re-calculated. The degradation rates are again compared with the QPCR based degradation rates and the microarray data adjusted appropriately again. This process is repeated until convergence is reached. This occurs very quickly taking less than 10 iterations for β to become less than 10^{-14}.

In the re-normalisation each microarray's data was rescaled so that the minimum degradation rate was approximately zero. Here the scale of each microarray is fixed but the levels are shifted up or down by a fixed amount. Furthermore, the amount each microarray has its values shifted is proportional to the time point the microarray represents. A rotational transformation is performed on the microarray data (in the log domain) so that the estimated degradation rates are in better agreement with the degradation rates estimated from QPCR (Figure 8.3).

Error Model

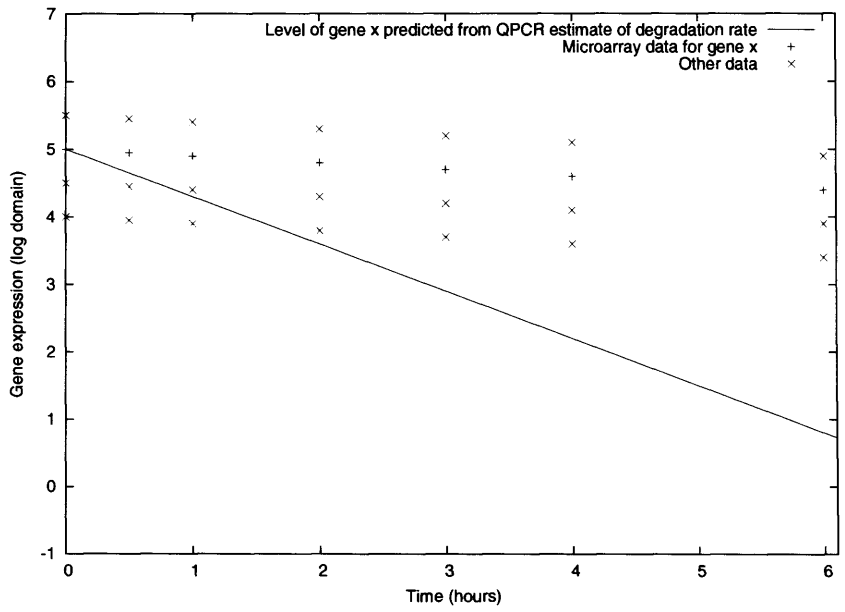
To use a line fitting routine some estimate of the level of measurement error is needed. Barenco *et al.* (2005) suggest that the level of error in a microarray measurement is principally associated with the amplitude of the expression signal and so measurement error can be expressed as a function of signal intensity. To estimate the relationship between error and signal a large data set of technical replicates is required. Here Affymetrix's spike-in microarray data set was used[4], which is made up of three technical replicates of 14 experiments produced using Affymetrix U133A microarrays. The measurement of gene j in experiment q and replicate a is denoted, $\bar{x}_{j,q,a}$. For each gene j and experiment a the mean and variance are calculated,

$$\bar{x}_{j,q} = \frac{1}{2}\sum_{a=1}^{3} \bar{x}_{j,q,a},$$

[3]For each gene there can be multiple probe sets on the Affymetrix array. All of the relevant probe sets are used in the comparison between the microarray and QPCR data apart from those that have two or more time points with a detection p-value greater that 0.4 or those that had a predicted degradation rate over twice the size of the degradation rates of the rest of the probe sets for that gene.

[4]`http://www.affymetrix.com/support/technical/sample_data/datasets.affx`

(a)



(b)

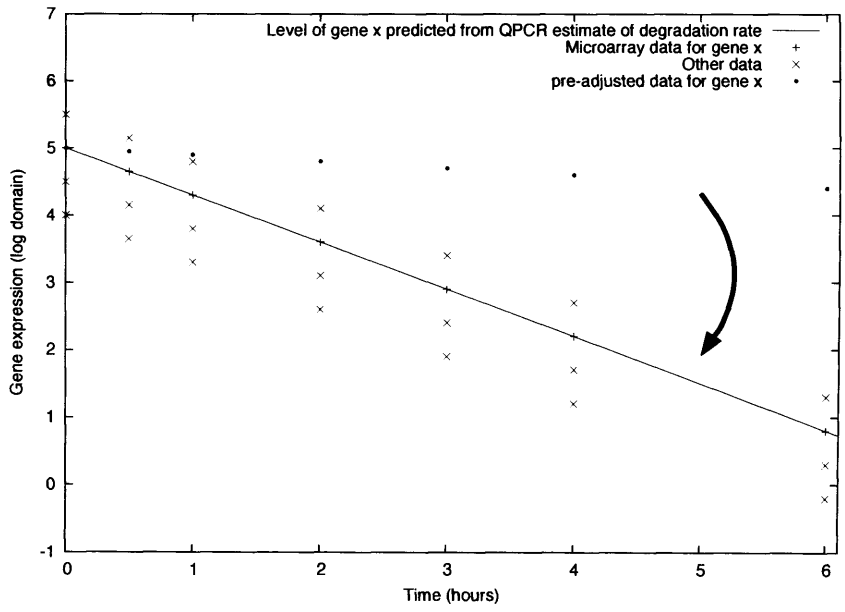


Figure 8.3: An idealised example of the process of anchoring microarray data using one known degradation rate from QPCR, with (a) data before adjustment and (b) data post adjustment. The degradation rate measured by QPCR for gene x is 0.7 and the pre-adjustment value for the microarray data is 0.1.

$$\bar{\sigma}_{j,q}^2 = \frac{1}{3-1}\sum_{a=1}^{3}(\bar{x}_{j,q,a} - \bar{x}_{j,q})^2.$$

$\bar{x}_{j,q}$ represents a *signal observation* with $\bar{\sigma}_{j,q}$ an estimate of its associated error (formally $\bar{\sigma}_{j,q}$ is the estimated standard deviation of a random variable with mean zero). The signal observations are then divided into thirty different classes, H_b, for $b = 1, \ldots, 30$. These classes are of equal width in the log domain. l_b is the lower log domain bound of class b and u_b is the upper log domain bound. The signal observation is a member of class H_b if $l_b \leq \ln(\bar{x}_{j,q}) < u_b$. u_b and l_b are defined as follows,

$$\begin{aligned} u_b &= l_{b+1}, \\ u_b - l_b &= \frac{1}{30}\left(\max_{j,q}[\ln(\bar{x}_{j,q})] - \min_{j,q}[\ln(\bar{x}_{j,q})]\right), \\ u_{30} &= \max_{j,q}[\ln(\bar{x}_{j,q})], \\ l_1 &= \min_{j,q}[\ln(\bar{x}_{j,q})]. \end{aligned}$$

For each class, one can associate a signal and an estimated error,

$$\begin{aligned} \ln(\hat{x}_b) &= \frac{1}{2}(u_b - l_b) \\ \sigma_b^2 &= \frac{1}{m_b}\sum_{j,q \in H_b} \bar{\sigma}_{j,q}^2, \end{aligned}$$

where m_b is the number of signal observations in class H_b. These pairs of values can then be used as a look up table to get an estimated error for a gene with expression level, $\hat{x}$. Linear interpolation and extrapolation (in the log domain) is used to find the errors outside the given points. Figure 8.4(a) shows a plot of the relationship between signal and error. The procedure used here is similar to that used in Barenco *et al.* (2005), but uses all 14 experiments instead of just one.

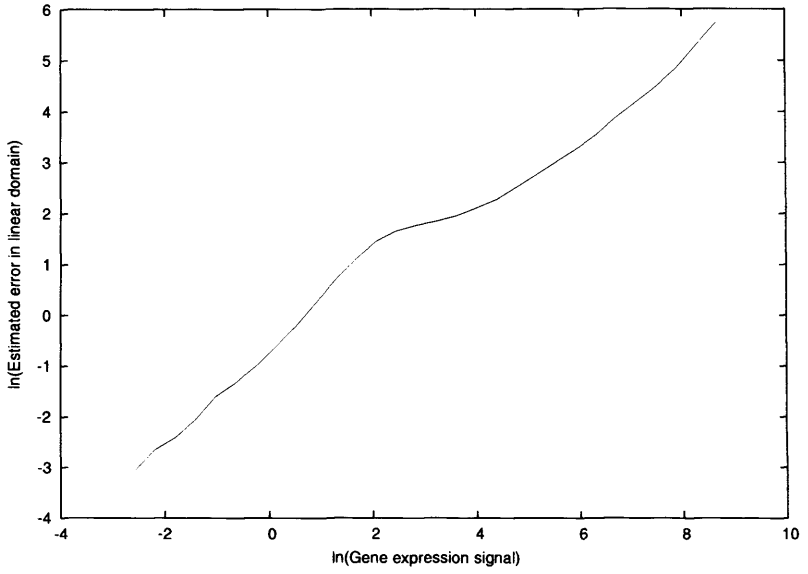
An estimated error is also required in the log domain because the linear regression is performed there. The spike-in data set was transferred into the log domain i.e. $\bar{x}_{j,q,a} = \ln\left(\bar{x}_{j,q,a}^{\text{orig}}\right)$ and then the above procedure was repeated (see Figure 8.4(b)).

Determination of degradation rates

The complete procedure including re-normalisation and QPCR anchoring was performed on the full microarray data set of 22277 genes. The re-normalisation reverses the normalisation performed by MAS5.0 by scaling each microarray so that the minimum degradation rate is approximately zero. The total scaling factors required for each microarray get smaller as time increases (Tables 8.2), this is expected because the mRNA decays away with time and so the average mRNA level (which MAS5.0 made equivalent) should decrease with time too.

After re-normalisation the microarray data is adjusted so that the estimated degra-

(a)



(b)

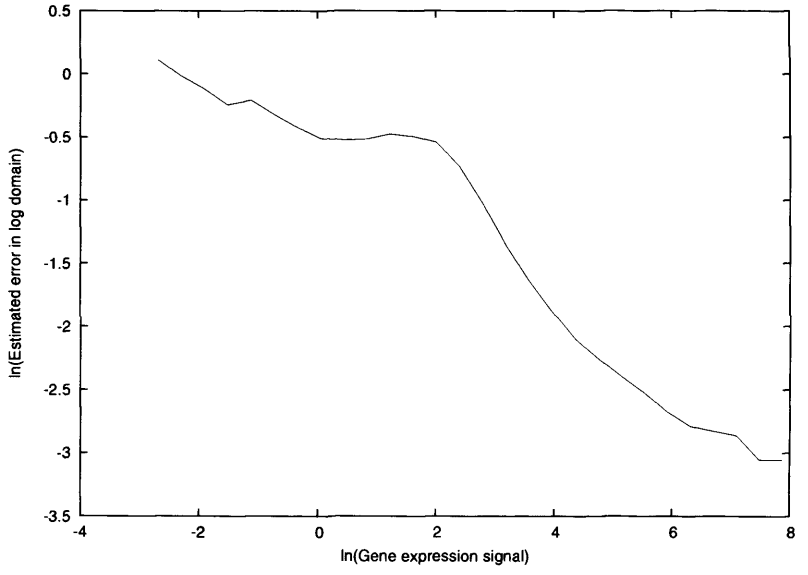


Figure 8.4: A plot to show the error associated with a particular gene expression signal in (a) the linear domain and (b) the log domain.

Table 8.2: The total scaling factors used on the degradation microarray time course data.

Time	0	0.5	1	2	3	4	6
Adjustment factor	1	0.893	0.753	0.6473	0.448	0.407	0.292

Table 8.3: The Affymetrix probe sets related to each gene used in the QPCR experiments.

Gene name	Associated Affymetrix tags
GADD45	203725_at
MDM2	217373_x_at
p21	202284_s_at
CD71	207332_s_at
HPRT1	202854_at
PGK1	200737_at, 200738_s_at, 217356_s_at, 221616_s_at

Table 8.4: A table to show the degradation rates estimated from the microarray data before and after the adjustment taking into account QPCR data.

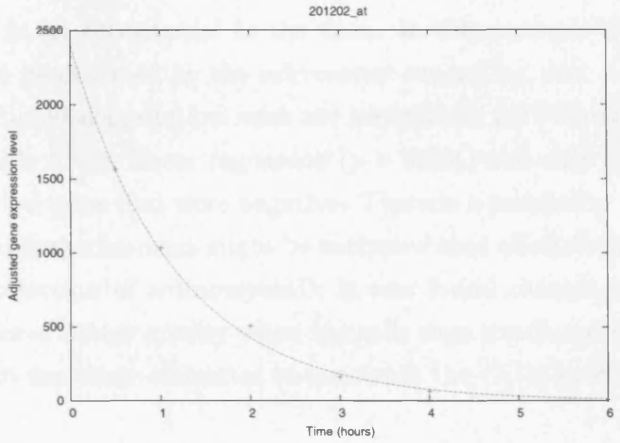
Affymetrix label	Gene name	Degradation rates		
		QPCR	Microarray before	Microarray after
203725_at	GADD45	1.68	0.841	1.27
217373_x_at	MDM2	0.253	0.168	0.591
202284_s_at	p21	0.904	0.703	1.10
207332_s_at	CD71	0.602	0.158	0.577
202854_at	HPRT1	0.522	0.130	0.540
200737_at	PGK1	0.531	0.112	0.527
200738_s_at	PGK1	0.531	0.128	0.536
217356_s_at	PGK1	0.531	0.127	0.535
221616_s_at	PGK1	0.531	0.193	0.592

dation rates are in the best possible agreement with the estimated degradation rates found from QPCR. Table 8.3 shows the Affymetrix probe sets used in the comparison between PCR and microarray data[5]. The total shift in the estimated degradation rates, β, was 0.409 (to 3 *s.f.*). The adjustment increases the degradation rates because before the adjustment the lowest degradation rates were set to approximately zero which is an underestimate. For the vast majority of genes this method brings the microarray degradation rate closer to the QPCR degradation rate, the only exception to this is MDM2 (Table 8.4). A possible reason could be that the associated probe set (217373_x_at) is expected to have a large amount of cross hybridisation and so this could be affecting the quality of the microarray degradation rate estimate.

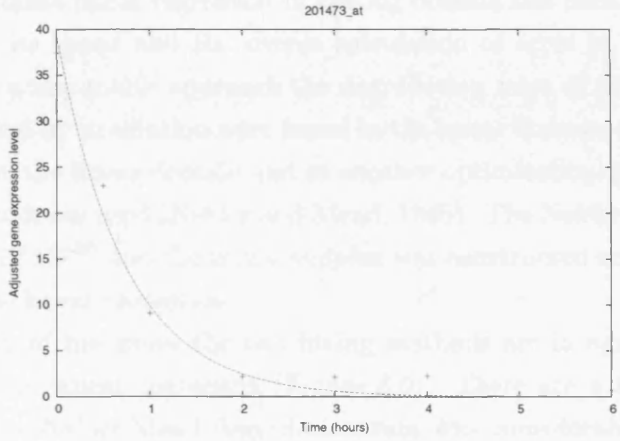
To check the procedure was working as expected a subset of 100 genes, chosen based on their activation by DNA damage, were examined in closer detail. This is important as the linear regression in particular has quite a few arbitrarily chosen constraints. The fit of the exponential (using the estimated degradation rate) to the adjusted microarray data was checked by eye. It was found that generally there was excellent agreement between the data and the exponentials (Figure 8.5(a)). Additionally, the algorithm seemed to

[5]The Affymetrix probe set labels provide some information about the amount of cross hybridisation that is expected to occur. "_at" indicates that the probe set is a perfect match and there are no other matches, "_s_at" indicates there is expected to be a small amount of cross hybridisation and "_x_at" indicates there is expected to be a large amount of cross-hybridisation.

(a)



(b)



(c)

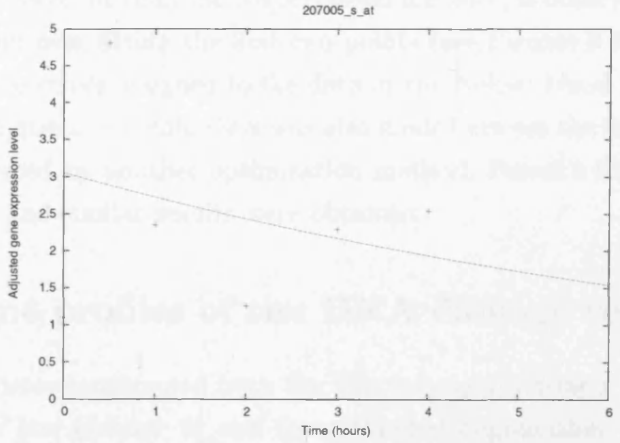


Figure 8.5: Examples of the degradation rates estimated and their fit to the data. (a) A standard data set (b) a data set with the saturation noise effect and (c) a badly fitting data set.

detect whether the tailing-off effect occurred or not (Figure 8.5(b)). Figure 8.5(c) is an example of a data set with a "bad" fit, it is not that the procedure is failing but that it is inappropriate to fit an exponential to the data. In this example the data is not at a high enough level to be detected by the microarray suggesting that it is stuck in noise.

Overall the estimated degradation rates are acceptable; for 97% of the probe sets the data it was reasonable to use linear regression ($p > 0.001$) and only 18 probe sets out of 22277 had degradation rates that were negative. There is a possibility that by irradiating the cells that apoptotic mechanisms might be activated that affect the rate of degradation even after the introduction of actinomycinD. It was found though, that the estimated degradation rates are of better quality when the cells were irradiated (appendix C.6) and so it is reasonable to use these estimates to construct the $G_g(t)$ profile.

Checking against exponential fit

Throughout this process linear regression in the log domain has been used as the fitting algorithm for both its speed and the simple calculation of error in the estimates. To test whether this is a reasonable approach the degradation rates of a subset of 100 genes known to be activated by irradiation were found in the linear domain. A linear regression is not appropriate in the linear domain and so another optimisation method, the Nelder-Mead simplex method was used (Nelder and Mead, 1965). The Nelder-Mead method was run to an accuracy of 10^{-10} and the initial simplex was constructed around the predicted parameters from the linear regression.

For the majority of the genes the two fitting methods are in agreement within the error predicted by the linear regression (Figure 8.6). There are a few outliers and in all of these cases the Nelder-Mead degradation rate was considerably higher than the linear regression value. On closer examination it appears that in these cases the linear fit actually produces a better fit than the Nelder-Mead method (as observed by eye) with the Nelder-Mead method over fitting the first two points (see Figures 8.7(a) – 8.7(d)). This may be caused by the errors assigned to the data or the Nelder-Mead routine may not be finding the global minima. A comparison was also made between the linear regression and the estimates produced by another optimisation method, Powell's direction set method (Press *et al.*, 2002) and similar results were obtained.

8.3 $G_g(t)$ time profiles of the DNA damage response

$G_g(t)$ time profiles were constructed from the 5Gy microarray time series data produced after DNA damage (see chapter 3) and the estimated degradation rates found above. The DNA damage response is a suitable system to analyse using this approach because a strong response is produced in a number of different transcription factors. The closer the transcription activator profile is to being constant the more difficult it will be to cluster the related $G_g(t)$ time profiles because it would be difficult to distinguish between noise

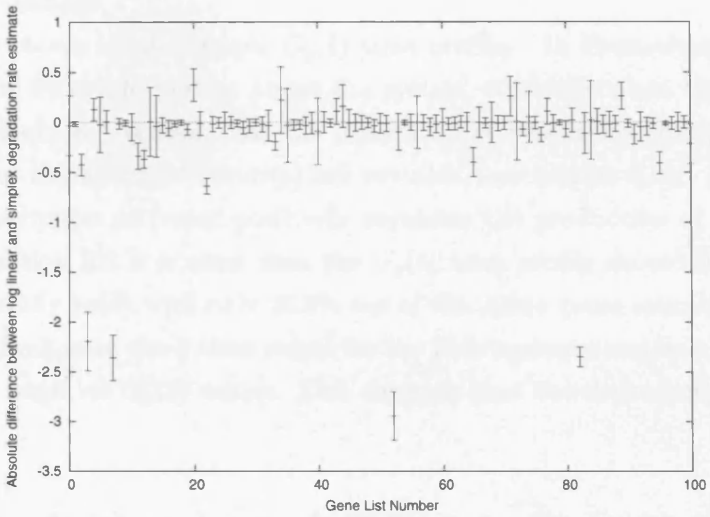


Figure 8.6: A plot showing the difference between the estimates of the degradation rate produced by the linear fit in the log domain and the exponential fit using the Nelder-Mead method. The error bars indicate the estimated error in the degradation rate according to the linear regression.

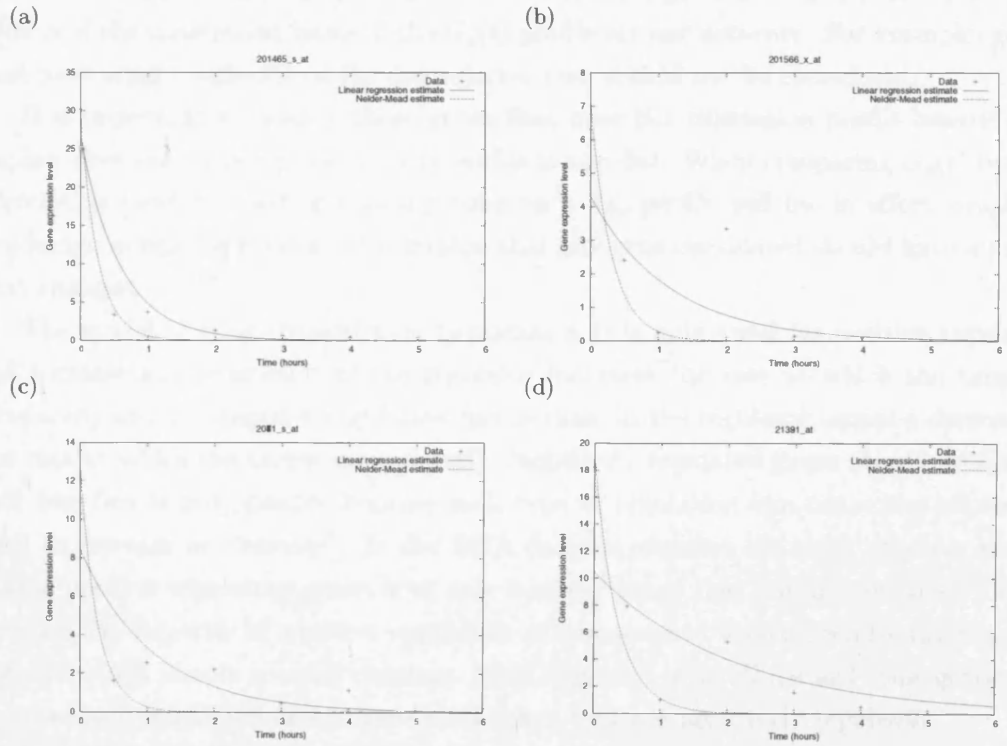


Figure 8.7: A plot showing the difference fits produced by the linear regression and the Nelder-Mead method for four genes (a) 201465_s_at, (b) 201566_x_at, (c) 208711_s_at and (d) 213931_at. These genes were chosen as they had large differences.

and the transcription activity profile. It is assumed that the system is in equilibrium prior to DNA damage.

Figure 8.8 shows some example $G_g(t)$ time profiles. In themselves the $G_g(t)$ time profiles provide little information about the system, it is only when the different $G_g(t)$ profiles are clustered together that the possibility of identifying groups of genes that share the same transcription factor(s) are revealed (see section 8.6). As it is assumed that the transcription activator positively regulates the production of the target gene, then from equation 8.2 it is clear that the $G_g(t)$ time profile should be positive. This condition generally holds with only 26.8% out of the 22284 genes containing one or more negative $G_g(t)$ values in the 8 time points for the first replicate and only 0.7% containing three or more negative $G_g(t)$ values. This suggests that the degradation rates used are reasonable.

8.4 Using $G_g(t)$ and a training set to find p53 targets

8.4.1 Motivation for the gene list

When using the $G_g(t)$ time profiles it is important not to consider profiles that are of poor quality so that accurate results are produced. Profiles should not be considered if the model of gene transcription (equation 8.1) is not a good description for a particular gene or if the constituent parts of the $G_g(t)$ profile are not accurate. For example, gene's that have a poor estimate of the degradation rate should not be considered.

It is important to remove those genes that have flat expression profile because this implies that the transcription activity profile is also flat. When comparing $G_g(t)$ profiles correlation must be used and so any noise on a flat profile will be, in effect, amplified producing unreliable results. This implies that any gene considered should have a profile that changes.

The model of gene transcription (equation 8.1) is only valid for positive regulation (an increase in the amount of the regulator increases the rate at which the target is produced) and not negative regulation (an increase in the regulator causes a decrease in the rate at which the target is produced). Negatively regulated genes should be filtered out, but this is not possible because each type of regulation can cause the expression level to increase or decrease[6]. In the DNA damage response the most effective way to isolate positive regulatory genes is to only consider genes that are up-regulated. This is because the majority of positive regulators will increase in activity whilst the negative regulators will remain roughly constant. This approach will still discard some genes that are positively regulated and include some genes that are negatively regulated.

[6]If a gene is controlled by positive regulation then if the regulator level goes up the gene expression level will increase and if the regulator goes down the gene expression level will decrease. Conversely, if a gene is negatively regulated then if the regulator level goes up the gene expression level will decrease and if the regulator goes down the gene expression level will increase.

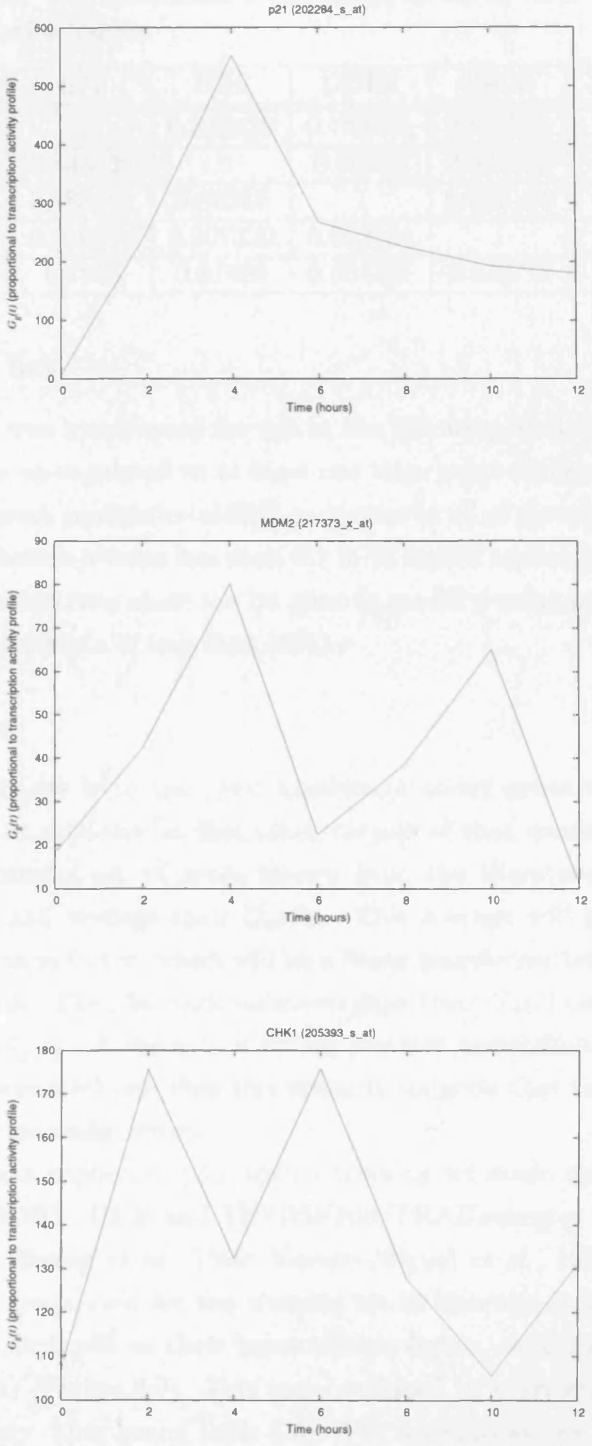


Figure 8.8: Three examples of $G_g(t)$ time profiles obtained from the first replicate of time series microarray data obtained after the cells were exposed to 5Gy of γ radiation. These are proportional to the activity that drives the mRNA levels of each gene and should not be confused with the expression levels.

Table 8.5: A correlation square to test whether the training set has sufficiently similar shaped G time courses. The correlation is calculated across all three replicates providing a more robust correlation value.

	p21	BIK	DDB2	PA26	TNFRSF10b
p21	1	0.838528	0.788661	0.810075	0.7855
BIK	0.838528	1	0.88365	0.907529	0.87488
DDB2	0.788661	0.88365	1	0.952104	0.834437
PA26	0.810075	0.907529	0.952104	1	0.830718
TNFRSF10b	0.7855	0.87488	0.834437	0.830718	1

8.4.2 The gene list

A list of 2355 genes was constructed for use in the following analysis. For a gene to be in the list it must be up-regulated at at least one time point (with respect to the initial time point with a z-score confidence of 85% or greater in all of the replicates), present (at one point with a detection p-value less than 0.1 in at least 2 replicates), and the estimate of the gene's degradation rate must not be poor (a model p-value greater than 0.01 or a relative error in the estimate of less than 10%).

8.4.3 Procedure

One application of $G_g(t)$ is to use prior knowledge about genes that share the same transcription factor or activator to find other targets of that transcription factor. The idea is to take a training set of genes known from the literature to share the same transcription factor and average their $G_g(t)$s. This average will give a *representative* $G_g(t)$ of the transcription factor, which will be a linear transformation of the transcription factor's activity profile. Then for each unknown gene their $G_g(t)$ can be correlated with this representative $G_g(t)$. If there is a strong positive correlation between the gene's $G_g(t)$ and the representative one then this strongly suggests that the gene is a target of the transcription factor under study.

This technique was applied to p53, with a training set made up of 5 significant p53 targets: p21, BIK, DDB2 , PA26 and TNFRSF10b/TRAILreceptor 2 (Bunz *et al.*, 1998; Marko *et al.*, 2003; Hwang *et al.*, 1999; Velasco-Miguel *et al.*, 1999; Wu *et al.*, 1997) (these are the same genes used for the training set in Barenco *et al.* (2005)). As these genes are known to have p53 as their transcription factor, then the genes should have the same shape $G_g(t)$ (Figure 8.9). This was confirmed by correlating each gene in the training set with every other gene (Table 8.5). The correlations are high for all genes in the training set with p21 being an outlier.

Before the representative G time course was constructed, each of the training set genes' G time courses was rescaled so that the minimum value of all three replicate G time series is zero and the maximum is one. This is so that each training set gene has an equal weight in the representative time course. This is important because in practice

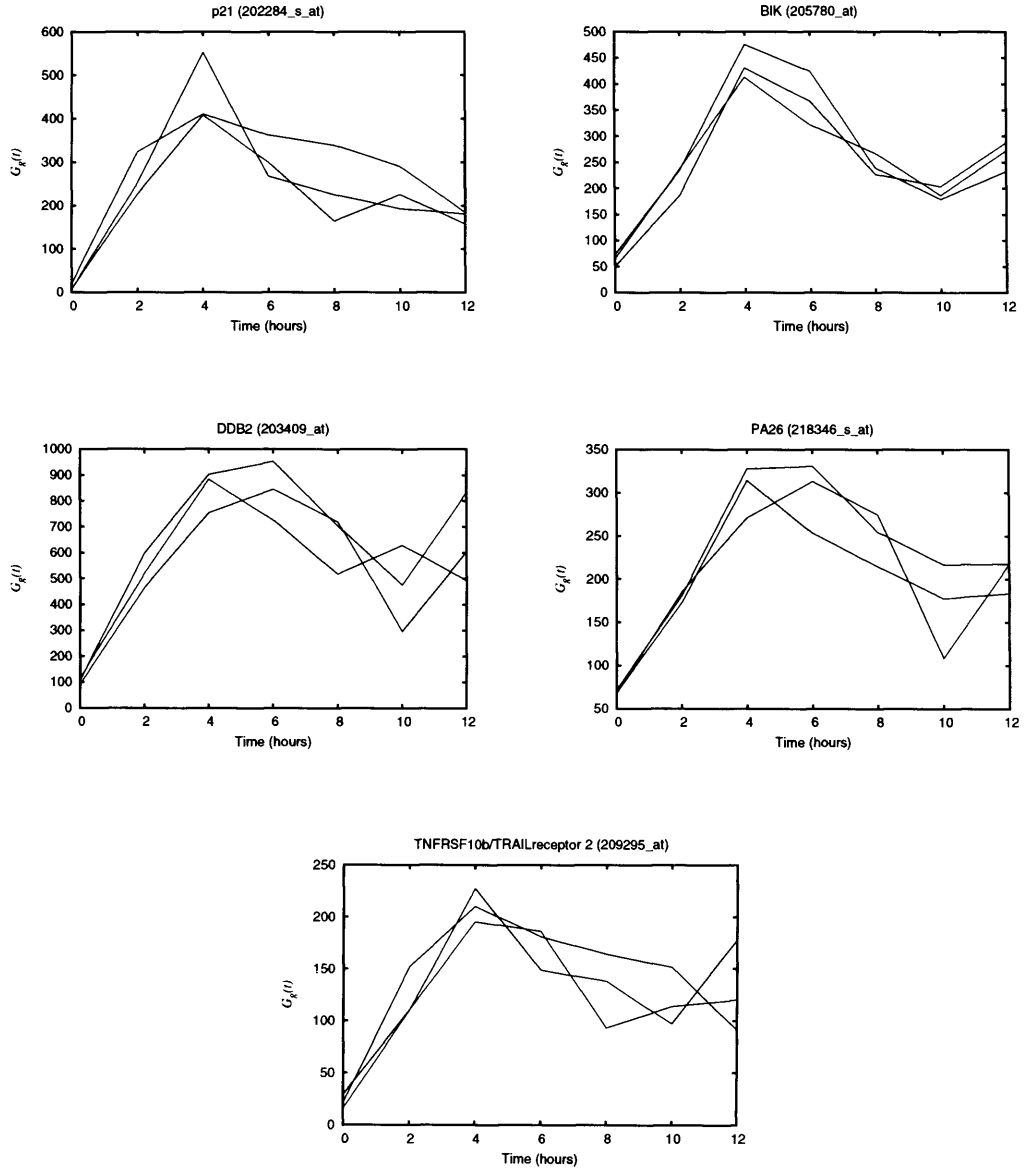


Figure 8.9: A plot showing the $G_g(t)$ profiles for all the members of the p53 training set. All three replicates are shown.

the correlation between the training set genes is not perfect; if no rescaling was done and one of the G time series had much larger values then the representative time course would have a shape closer to that gene than the others. As the rescaling is a linear transformation of the data, it will not affect the correlation it will have with other genes. For all three replicates the representative G profile rapidly rises to a peak at 4 hours and then decays away at a slower rate (Figure 8.10(a)). Even though the three replicates follow broadly the same shape there are slight differences, such as both replicate 2 and 3 starting to rise again between 10 and 12 hours. The replicates are biological replicates so differences are expected. The activity profile of p53 was measured by the quantification of a Western blot that used an antibody that detects p53 phosphorylated at serine 15 (Banin *et al.*, 1998), an accepted measure of p53 activity (Figure 8.10(b)). The two profiles are similar but not the same, they both share the rapid early response, but the later decline predicted by the representative G time course is much more rapid than seen in the Western blot. A possible reason for this discrepancy is that even though the concentration of phosphorylated p53 is correct, it is not functionally active as it has been regulated by some other mechanism, such as the de-localisation of p53 from the nucleus (Li *et al.*, 2003). This is a prediction that could be experimentally verified. On balance, the G time profile does provide an accurate representation of the activity of p53 during the DNA damage response. Encouragingly, the representative G time series is virtually identical to that predicted through parameter estimation by Barenco *et al.* (2005). Therefore, the estimates for the degradation rates found by Barenco *et al.* (2005) are similar to the measured degradation rates. This suggests that the methodology for measuring the degradation rates is reasonably accurate and also that the best fit of the model produces appropriate parameter estimates i.e. the model is valid for p53 targets.

8.4.4 p53 verification experiment

A verification experiment was performed using small interfering RNA (siRNA). siRNA are a class of 20-25 nucleotide-long RNA molecules that bind with specific RNA transcripts, forming double-stranded mRNA. This is targeted for degradation and so the introduction of siRNA for a particular gene effectively stops its expression. siRNAp53 is siRNA that prevents the expression of p53. A group of MOLT4 cells were split into two groups with one group being transfected through electroporation with a vector expressing siRNAp53 and the other with a vector-only control. Both were also transfected with a marker plasmid carrying a copy of a gene encoding GFP which was used to FACS sort GFP expressing cells to a purity of greater than 98%. In one group the p53 protein will be considerably depleted and in the other p53 will act as normal.

After 48 hours, each culture was split into two samples, one sample that was irradiated with 5Gy of gamma rays and the other that was mock irradiated. All four samples were then incubated at 37 °C for four hours. RNA was prepared from all four samples and used in microarray experiments. Some of each sample were kept for QPCR or protein

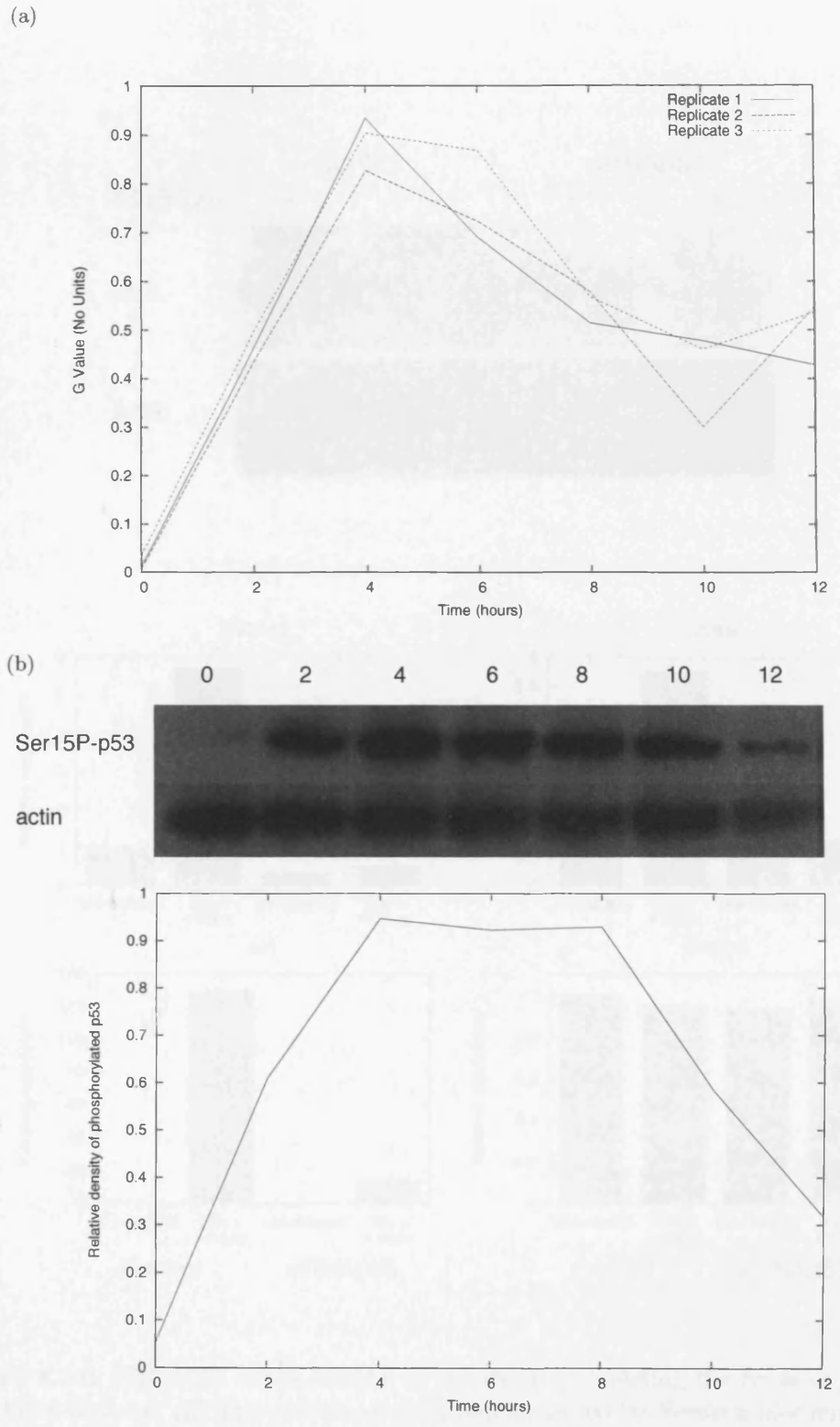


Figure 8.10: A plot showing (a) the representative G time course for each replicate produced from a training set of p53 genes and (b) the relative density of active p53 protein (phosphorylated p53) in replicate one of the experiment measured by quantification of Western blot data. Western blot performed by Daniela Tomescu.

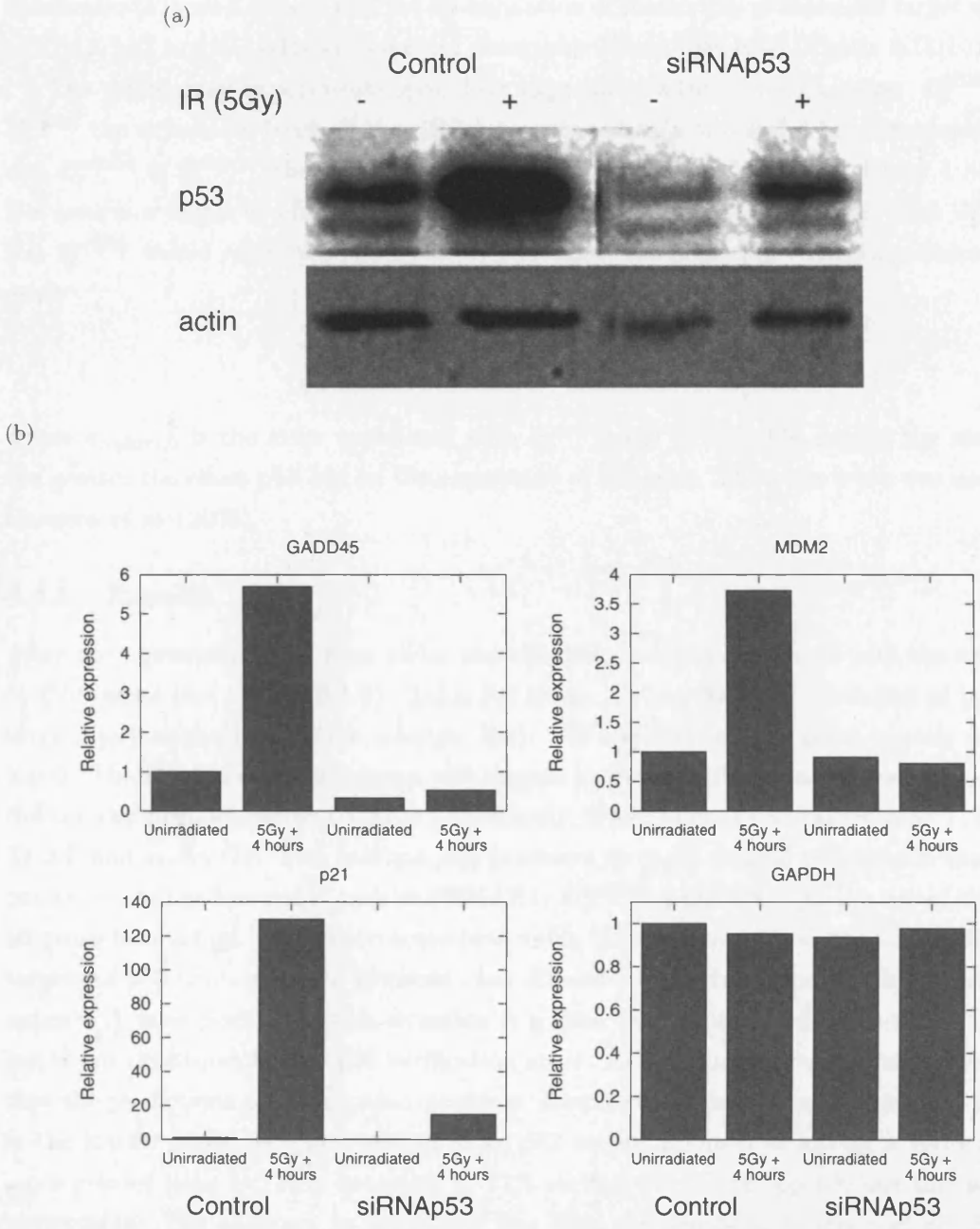


Figure 8.11: siRNAp53 was successful in significantly depleting the levels of p53 even after DNA damage. (a) The protein levels of p53 measured by Western blot and (b) the mRNA levels of three known p53 targets (GADD45, MDM2 & p21) and one standard (GAPDH) measured by QPCR. Western blots and QPCR performed by Daniela Tomescu.

analysis. It was confirmed that the siRNAp53 had successfully been transfected and was active by observing that the concentration of p53 protein was significantly depleted after irradiation (Figure 8.11(a)), and the up-regulation of transcripts of three p53 target genes (MDM2, p21 and GADD45α) were not detectable above base level (Figure 8.11(b)).

The microarray experiments gave four expression values for each gene: $\hat{x}_0^{\text{siRNA}}$ & $\hat{x}_4^{\text{siRNA}}$ the expression levels of the siRNA transfected cells at 0 and 4 hours respectively, and $\hat{x}_0^{\text{control}}$ & $\hat{x}_4^{\text{control}}$, the expression levels of the vector only control at 0 and 4 hours. If a gene is a target of p53 one would expect that $\hat{x}_4^{\text{control}}$ would be greater than $\hat{x}_0^{\text{control}}$ but $\hat{x}_0^{\text{siRNA}}$ would roughly equal $\hat{x}_4^{\text{siRNA}}$. Therefore, the following verification score was used,

$$V = \frac{\left(\hat{x}_4^{\text{control}} - \hat{x}_0^{\text{control}}\right) - \left(\hat{x}_4^{\text{siRNA}} - \hat{x}_0^{\text{siRNA}}\right)}{\sqrt{\sigma^2_{\text{control},0} + \sigma^2_{\text{control},4} + \sigma^2_{\text{siRNA},0} + \sigma^2_{\text{siRNA},4}}},$$

where $\sigma_{\text{control},0}$ is the error associated with $\hat{x}_0^{\text{control}}$ and so on. The higher the score is the greater the effect p53 has on the expression of the gene. The same score was used in Barenco *et al.* (2005).

8.4.5 Results

After the representative G time series was calculated, it was correlated with the sub-set of 2355 genes (see section 8.4.2). Table 8.6 shows the top 50 genes predicted to be p53 targets (see appendix D.2 for a longer list). All the training set genes appear in the top 6. There are many well known p53 targets in the top 100 including members from the tumour necrosis factor receptor superfamily, GADD45α, TP53 target gene 1, SNK, TRAF and cyclin G1. This method also produces strongly verified p53 targets that are not known in the literature, such as PRKAB1, ASCC3L1 and P45. At the top of the list all genes have a high verification score confirming that this analysis method is predicting targets of p53 (Figure 8.12). Between rank 20 and 30, the first apparent false positives appear. A false positive in this situation is a gene that is predicted to be a p53 target but is not confirmed by the p53 verification score. As the rank increases the probability that the predictions are inaccurate increases. Despite this the results are generally good, in the top 50 genes, 84% are verified to be p53 targets (defined as having a verification score greater than 1). This decreases to 71% verified out of the top 100 but this is still respectable. The accuracy is marginally less than the results in Barenco *et al.* (2005) where 90% of the top 74 genes were verified. This is mainly because the method used by Barenco *et al.* (2005) filters out data that cannot be described by the model (which is impossible by the $G_{g(t)}$ approach) and then ranks the remaining data by each gene's sensitivity to p53.

Interestingly the verification experiment is not infallible. FBXW7 is predicted at rank 24 to be a likely p53 target but the verification score is -0.541. Mao *et al.* (2004) found that FBXW7 was in fact p53-dependent. So some apparent false positives suggested by

Table 8.6: A list of the top 50 genes predicted to be targets of p53. The shaded entries are the training set gene. Citations: [1] Velasco-Miguel *et al.* (1999), [2] Marko *et al.* (2003), [3] Hwang *et al.* (1999), [4] Wu *et al.* (1997), [5] Bunz *et al.* (1998), [6] Park and Nakamura (2005), [7] Obad *et al.* (2004), [8] Bates *et al.* (1996), [9] Li *et al.* (2004a), [10] Varmeh-Ziaie *et al.* (1997), [11] Smith *et al.* (1994), [12] Fiscella *et al.* (1997), [13] Mao *et al.* (2004), [14] Liu and Chen (2002) & [15] Amundson *et al.* (2002).

Affymetrix code	Description	Correlation Value	Verification Score	p53 citation
218346_s_at	p53 regulated PA[illegible] nuclear protein	0.963	3.90	[1]
205780_at	BCL2-interacting killer (BIK)	0.960	6.57	[2]
203409_at	damage-specific DNA binding protein 2 (DDB2)	0.954	10.7	[3]
209295_at	tumour necrosis factor receptor superfamily, member 10b	0.921	6.52	[4]
218627_at	hypothetical protein FLJ11259	0.896	3.56	–
[illegible]_s_at	cyclin-dependent kinase inhibitor 1A (p21)	[illegible]	[illegible]	[5]
201834_at	protein kinase, AMP-activated, β 1 non-catalytic subunit (PRKAB1)	0.888	6.30	–
204674_at	lymphoid-restricted membrane protein	0.865	3.40	–
218403_at	p53-inducible cell-survival factor (P53CSV)	0.852	7.75	[6]
212371_at	CGI-146 protein	0.850	2.61	–
213293_s_at	tripartite motif-containing 22 (TRIM22,STAF50)	0.847	6.07	[7]
208796_s_at	cyclin G1	0.844	5.18	[8]
215719_x_at	tumour necrosis factor receptor superfamily, member 6 (FAS)	0.830	8.11	[9]
219628_at	p53 target zinc finger protein (WIG1)	0.819	3.70	[10]
212815_at	activating signal cointegrator 1 complex subunit 3 (ASCC3L1)	0.817	5.93	–
216252_x_at	tumour necrosis factor receptor superfamily, member 6 (FAS)	0.814	4.54	[9]
205692_s_at	CD38 antigen (P45)	0.798	9.02	–
203725_at	growth arrest and DNA-damage-inducible, alpha (GADD45α)	0.798	11.0	[11]
204566_at	protein phosphatase 1D magnesium-dependent, δ isoform (PPM1D)	0.798	6.05	[12]
212430_at	RNA-binding region containing 1 (RNPC1)	0.772	2.33	–
213038_at	IBR domain containing 3	0.769	2.79	–
219361_s_at	hypothetical protein FLJ12484	0.755	5.43	–
218751_s_at	F-box and WD-40 domain protein 7 (FBXW7)	0.746	-0.541	[13]
207813_s_at	ferredoxin reductase	0.745	7.72	[14]
201835_s_at	protein kinase, AMP-activated, β 1 non-catalytic subunit (PRKAB1)	0.744	5.92	–
202726_at	ligase I, DNA, ATP-dependent (LIG1)	0.739	2.69	–
218007_s_at	ribosomal protein S27-like (RPS27L)	0.736	9.36	–
218031_s_at	checkpoint suppressor 1 (CHES1)	0.734	1.15	–
207426_s_at	tumour necrosis factor (ligand) superfamily, member 4 (TNFSF4)	0.723	5.26	–
202181_at	KIAA0247	0.720	2.22	–
203562_at	fasciculation and elongation protein zeta 1 (FEZ1)	0.718	-1.86	–
218527_at	aprataxin (APTX)	0.707	-2.32	–
209375_at	xeroderma pigmentosum, complementation group C (XPC)	0.703	5.80	[15]
207616_s_at	TRAF family member-associated NFKB activator (TANK)	0.703	-0.617	–
200730_s_at	protein tyrosine phosphatase type IVA, member 1 (PTP4A1)	0.700	4.45	–
211318_s_at	RAE1 homolog	0.695	3.44	–
217743_s_at	transmembrane protein 30A	0.692	2.33	–
204780_s_at	tumour necrosis factor receptor superfamily, member 6 (FAS)	0.691	7.78	[9]
205349_at	guanine nucleotide binding protein (G protein), α 15 (GNA15)	0.689	5.15	–
201093_x_at	succinate dehydrogenase complex, subunit A, flavoprotein (SDHA)	0.686	0.716	–
218288_s_at	hypothetical protein MDS025	0.684	2.38	–
214771_x_at	myosin phosphatase-Rho interacting protein (M-RIP)	0.679	-0.935	–
203846_at	tripartite motif-containing 32 (TRIM32)	0.676	0.862	–
35974_at	lymphoid-restricted membrane protein (Jaw1)	0.673	3.69	–
36564_at	IBR domain containing 3	0.672	3.19	–
219815_at	galactose 3-O-sulfotransferase	0.671	3.12	–
205531_s_at	glutaminase 2 (GLS2)	0.663	2.52	–
219099_at	chromosome 12 open reading frame 5	0.660	6.49	–
204060_s_at	protein kinase, X-linked	0.653	2.28	–
219627_at	hypothetical protein FLJ12700	0.652	0.801	–

the verification experiment may still be p53 targets. This problem occurs because the verification experiment only checks whether the gene is up-regulated at one time point, 4 hours, therefore it is bound to miss those genes that are late risers. Therefore, if anything the verification experiment under-estimates the performance of this procedure.

Ideally, one would have two distinct groups, one that had a high correlation to the representative G time profile and a high verification score and another with a low correlation and low verification score. This does not occur, for the following reasons:

1. The G time course is inaccurate to a certain degree because of the potentially large errors in the constituent parts: the estimated degradation rate, the estimated gradient and the microarray measurements. Therefore the G time course may match the representative profile when it should not. As one moves down the list

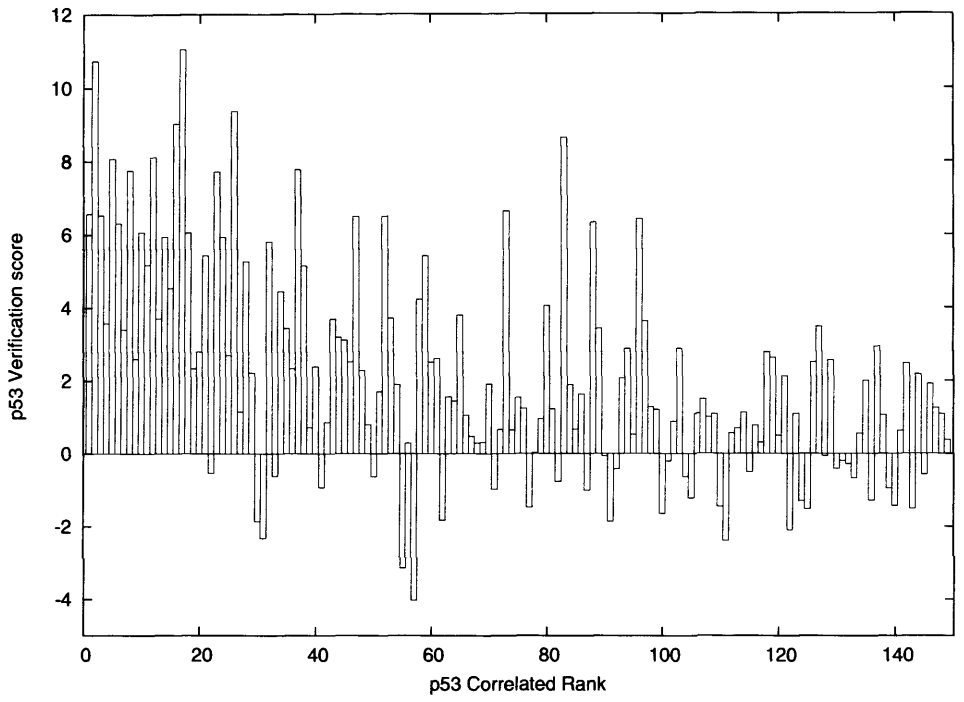


Figure 8.12: A plot showing how the verification score varies with rank in a list of 150 most likely p53 targets.

the correlation decreases, so the transcriptional activator profile becomes further away from the representative p53 G time series and so the probability that the G time series (within error) matches the true p53 activity profile is reduced. This explains why at the lower end of the list the accuracy diminishes.

2. The transcription of a gene might not be well described by the simple transcription model. The gene's mRNA level might be negatively regulated, regulated by two or more separate transcription factors acting in co-operation or the transcriptional response might be highly non-linear. It is more likely that this effect will cause a p53 target to have a poor correlation with the representative profile rather than the opposite but it is possible. This technique can not judge whether the model fits the data or not because it is correlation based because it is correlation based. The method just assumes the model does fit, unlike the Barenco *et al.* technique.

3. The possibility that two or more transcription activators have similar profiles. This technique can not distinguish between two transcription activators that have the same profile, so if there are two that are similar, a target of one will strongly correlate with the other. This would produce an inaccurate prediction.

8.4.6 A comparison with results from the gene expression profile

The aim of the $G_g(t)$ training set and correlation procedure described above is to improve the accuracy of predictions or at least detect different targets to alternative methods. The

correlation procedure uses an estimation of hidden factors (the shape of the transcription activator profile) to predict p53 targets. This procedure worked well with 84% of the top 50 predicted p53 targets being verified. To assess the significance of the training set and correlation procedure and using $G_g(t)$ data two alternative methods were analysed: clustering the expression profiles using k-means and following the correlation procedure but using expression profiles.

The traditional approach to finding genes that share the same function and are driven by the same activity is to use a clustering method (see appendix A.3.4). K-means clustering (Sherlock, 2000) was applied to the filtered 5Gy raw gene expression data, with the desired number of clusters set to eight as in Barenco *et al.* (2005). It was found that the top 50 p53 gene targets predicted by the $G_g(t)$ correlation procedure (Table 8.6) were distributed between 6 of the 8 clusters (Figure 8.13); the majority (68%) of predictions were in one cluster and a substantial amount was in another (20%), with the few remaining being spread among four other clusters. The apparent false positives found from the verification experiment were evenly split between the two major clusters. Up to 20% of the top 50 p53 targets predicted would not have been detected if clustering was used.

The correlation procedure set out in section 8.4.3 was repeated using gene expression profiles instead of the $G_g(t)$s. Each gene's expression profile was correlated with a representative gene expression profile instead of a representative activity profile. Overall there is a small but significant improvement in the accuracy of the targets if the G time profile is used (Figure 8.14); apart from between rank 40 and 53, the percentage of verified targets when G time profiles are used is higher than the gene expression profiles. There is a large overlap between the two sets of predictions; out of the top 50 predictions made, 75% are shared by both sets of data and out of the top 100 88% of predictions are common.

The $G_g(t)$ training set and correlation procedure (section 8.4.3) makes a significant improvement over clustering methods and a minor improvement over using the correlation procedure with the gene expression profiles. The difference between the gene expression profile and $G_g(t)$ depends on the degradation rate, therefore the improvement made by using G time profiles will depend on the distribution of the degradation rates (section 8.1.4), but it will also depend on the quality of the measurements used to construct $G_g(t)$. In the clustering the majority of the top 50 predictions occur in one cluster and in the correlation of gene expression profiles there was a large overlap in the predicted targets. This suggests that a large proportion of p53 targets share similar mRNA degradation rates.

The degradation rates of the positively verified genes that occur in the top 100 list produced from the G time profiles but do not occur in the corresponding list produced from the gene expression data are spread over a range of values (Table 8.7), but compared to the average degradation rate (0.675 hr^{-1}) they generally are at extreme values. When the degradation rates are high the gene's expression level peaks early compared to the

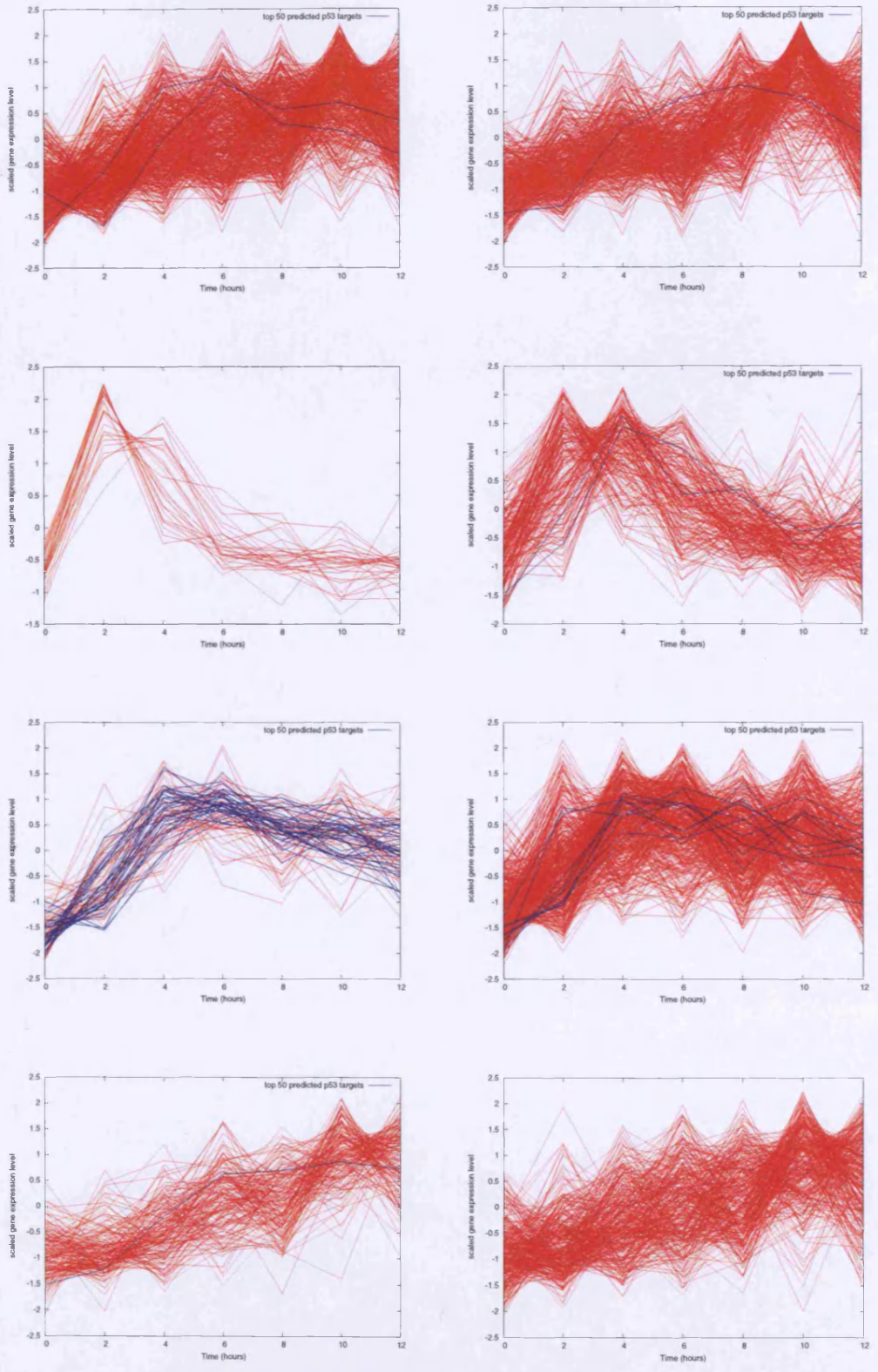


Figure 8.13: The result of k-means clustering on the gene expression profiles. The blue lines are the top 50 predicted p53 targets suggested by $G_g(t)$ correlation.

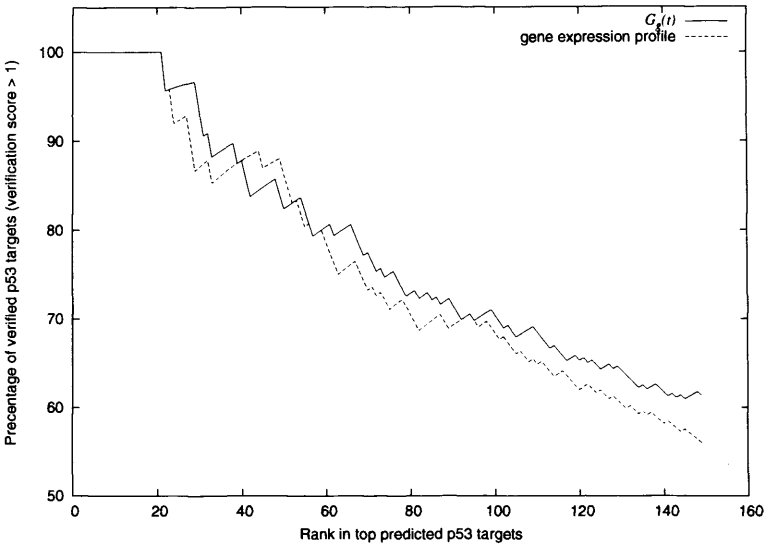


Figure 8.14: A plot showing how the performance varies with rank in a list of 150 most likely p53 targets for both G time profile data and gene expression data.

Table 8.7: Genes that are present in the top 100 list of predicted p53 targets from the G time series that do not occur in the top 100 list predicted from the gene expression profile and are verified to be p53 targets (verification score > 1).

Gene Name	Description	$G_g(t)$ Rank	Expression Profile Rank	Verification Score	Deg. Rate (hr^{-1})
205349_at	Guanine nucleotide binding protein α15	39	107	5.15	0.884
36564_at	IBR domain containing 3	45	164	3.19	1.31
209917_s_at	TP53 target gene 1	81	123	4.05	0.540
202693_s_at	serine/threonine kinase 17a (apoptosis-inducing)	84	180	8.63	1.40
217804_s_at	Interleukin enhancer binding factor 3 (ILF3)	94	193	2.07	0.611
204683_at	Intercellular adhesion molecule 2 (ICAM2)	99	235	1.28	0.806
204391_x_at	Transcriptional intermediary factor 1 (TIF1)	100	142	1.21	1.18

representative expression profile (Figure 8.15(a) and 8.15(b)), whilst if they are low the expression peaks later (Figure 8.15(c)). The $G_g(t)$ profiles are significantly closer to the corresponding representative profile than the gene expression profiles (Figure 8.15). When searching for targets a cut off has to be made somewhere, so a different set of targets would be gained by using $G_g(t)$. A possible reason for the large overlap is that the two representative profiles are fairly similar (Figure 8.15). In other systems where the range of degradation rates of the targets is larger, the improvement by using the G time profile data would be more significant.

The improvement of using $G_g(t)$ in the correlation occurs despite the bias against the $G_g(t)$ data due to the verification experiment being based around the level at four hours. Expression profiles that show a fast response are more likely to be verified and these are likely to have similar degradation rates. The full potential of using G time profiles is not being exploited. It is likely that if a later time point or a series of time points were used that the number of verified p53 targets by $G_g(t)$ profile data would significantly increase.

Apart from the performance there is an additional benefit to using a training set and

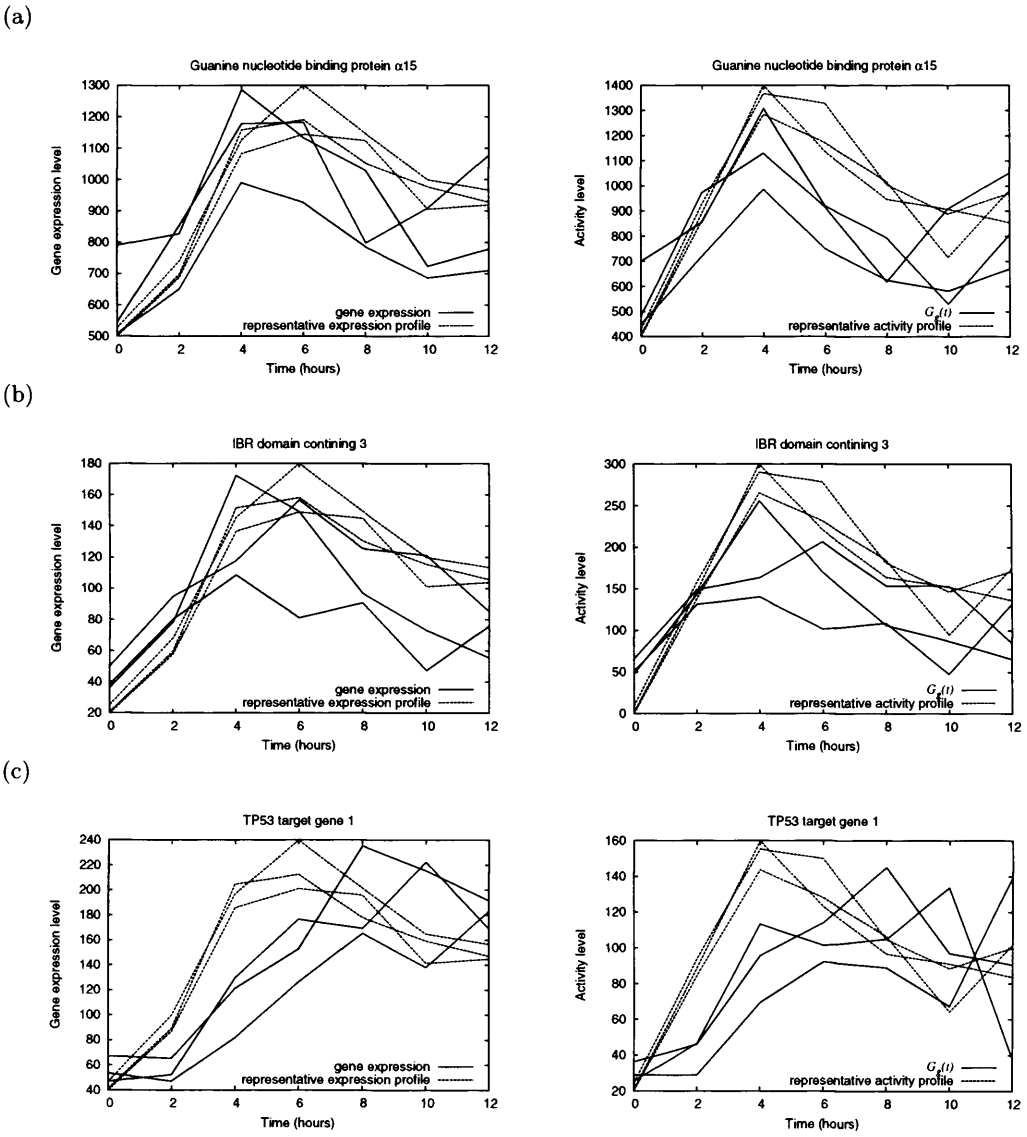


Figure 8.15: Plots comparing the gene expression and G profile with their associated p53 representative profiles for (a) Guanine nucleotide binding protein $\alpha 15$, (b) IBR domain containing 3, and (c) TP53 target gene 1. The representative profiles are rescaled to match the gene levels.

correlation procedure rather than a clustering method. In the correlation method a list of predicted targets is ranked given a quantitative assessment of the likelihood that a gene is a target. This is not possible in clustering, where a gene is either a member of a cluster or not, and one has to choose the cluster that is assigned to a particular transcription activity profile. This makes finding novel targets little better than guess work. Also by using the G time profile instead of the gene expression profile the shape of the transcription activity of p53 is gained without performing any additional experiments, which can be of considerable interest in its own right.

8.5 Clustering of the G time profiles

It would be useful if information could be gained about a system without knowledge about a transcription factor and its targets. It would be beneficial to also gain a description of the whole system, not just a single part (using the training set approach there is no way of telling whether a gene's $G_g(t)$ profile is a better correlation with a different transcription activator profile). In particular it would be interesting to know the number of significant activity profiles at work in a particular response and to see what the shape of these profiles are. For these reasons it is appropriate to use clustering techniques on the G time profile data. Clustering will divide the genes into groups that have strongly correlated G time profiles. If it is assumed that the simple transcription model is obeyed and that the G time profile is accurate enough, then the clusters of genes will correspond to genes that are transcribed by the same transcription activator profile (and probably the same activator). The average G time profile of a cluster will have the same shape as the activator that is transcribing the members of that cluster. The distance metric used must be Pearson correlation.

One of the problems with clustering is that generally each gene must be associated with one and only one cluster, this means that even though a gene's G profile might not correlate well with any of the clusters it will still be placed in the best cluster. By choosing the list of genes to cluster carefully one can minimise this effect. If the transcription of a gene is not well described by equation 8.1 then the corresponding G time profile will not correspond to the transcriptional activator of that gene so any cluster it is placed in will be inaccurate. It is likely that if this is the case, then the G profile will not correlate well with any of the principal activator profile shapes, but this can not be guaranteed.

8.5.1 Clustering Validation

Many standard clustering techniques require the number of clusters to be fixed either before the clustering is performed or after, when interpreting the results. The number of clusters chosen can have a large effect on the results and conclusions that are made. Therefore, it is important that if there is no clear idea of the number of clusters that some

technique is used to find the optimal number, this is called clustering validation (Halkidi *et al.*, 2001).

In preparation for clustering validation all the $G_g(t)$ time profile data was rescaled so that the minimum point was at 0 and the maximum point was at 1. This does not affect the correlation, but allows the finding of the cluster centres easily. Using k-means clustering (see appendix A.3.5) the list of 2355 genes were divided into clusters with the total number of clusters set between two and twelve using GeneSpring. A good cluster is normally defined as one that has small internal variation and is significantly separated from the other clusters. The Davies-Bouldin index is one common clustering validation index (Davies and Bouldin, 1979). Let c_i be the ith cluster of n_c clusters. The internal scatter of cluster i is defined as,

$$s_i = \frac{1}{n_i} \sum_{j \in c_i}^{n_i} \text{dist}(G_j(t), \bar{G}_i(t)),$$

where n_i is the number of members of cluster c_i, $\bar{G}_i(t)$ is the centre of cluster c_i and $\text{dist}(x, y)$ is the distance between two profiles x and y (in this case $1- < x, y >$). The dissimilarity score, d_{ij}, between clusters c_i and c_j is the distance between the two centres,

$$d_{ij} = \text{dist}(\bar{G}_i(t), \bar{G}_j(t)).$$

The Davies-Bouldin index, DB_{n_c}, is then calculated as follows,

$$\begin{aligned} R_{ij} &= \frac{s_i + s_j}{d_{ij}}, \\ R_i &= \max_{j=1,\ldots,n_c, j \neq i} R_{ij}, \\ DB_{n_c} &= \frac{1}{n_c} \sum_{i=1}^{n_c} R_i. \end{aligned}$$

The number of clusters that give the smallest Davies-Bouldin index is the optimal. The Dunn index, D_{n_c}, takes a similar approach but with different definitions of dissimilarity and internal scatter (Dunn, 1974). The dissimilarity is defined as,

$$d_{ij} = \min_{x \in c_i, y \in c_j} \text{dist}(G_x(t), G_y(t)),$$

and the internal scatter is defined as the diameter of the cluster,

$$s_i = \max_{x,y \in c_i} \text{dist}(G_x(t), G_y(t)).$$

Finally,

$$D_{n_c} = \min_{i=1,\ldots,n_c} \left\{ \min_{j=i+1,\ldots,n_c} \left(\frac{d_{ij}}{\max_{k=1,\ldots,n_c} s_k} \right) \right\}.$$

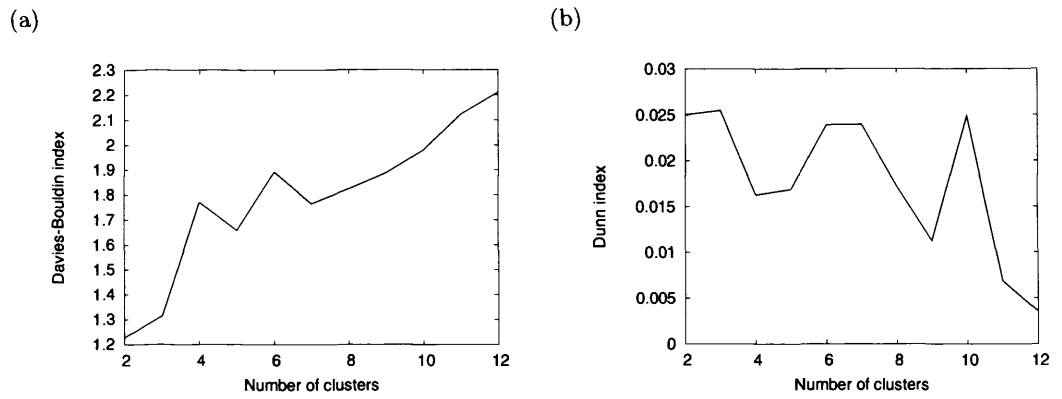


Figure 8.16: Cluster validation plots. A plot to show how (a) the Davies-Bouldin and (b) the Dunn index vary with the number of clusters using k-means clustering on $G_g(t)$ profile data. For the Davies-Bouldin index the smaller the index the better and for the Dunn index larger index is better.

For the Dunn index, the number of clusters that gives the largest index is the optimal.

These two indexes vary with the number of clusters produced by k-means (Figure 8.16(a)). The Davies-Bouldin index suggests two clusters are optimal and gets worse as the number of clusters increases (Figure 8.16(a)). As the optimal is two clusters and the Davies-Bouldin index cannot be used on one cluster this might suggest that there is no distinct clusters in the data. For the Dunn index, the optimal cluster size is three (Figure 8.16(b)), but the scores for clusters of 2, 3, 6, 7 and 10 are very similar. It is difficult to be confident about which number of clusters is optimal.

Francesca Buffa used the balanced iterative reducing and clustering hierarchies method on the $G_g(t)$ data (Buffa *et al.*, 2004; Chiu *et al.*, 2001). This method is based on the hierarchical method but allows optimisation of the number of clusters, it uses a bayesian information criterion. This method could not find any statistically significant clusters. This method is particularly good at picking out isolated clusters, so it is an indication that even if a gene is a member of one cluster its G profile could be reasonably well correlated with other clusters.

Three different cluster validation methods were applied to the $G_g(t)$ data and the results suggest that there is no definitive optimal number of clusters. This is indicative that the use of traditional clustering on $G_g(t)$ data to gain information about the global response to DNA damage is not appropriate. There are probably groups of genes that share very similar G profiles but this information is being lost in the whole data set due to "noisy" $G_g(t)s$. There are numerous sources of noise including the microarray data, the estimation of the degradation rates, the estimation of the rate of change in the data, and probably most importantly the model of gene transcription. The problem is that clustering techniques have to take into account $G_g(t)$ patterns that are dominated by noise that will correlate reasonable well with a large number of other genes including the tight groups. This means that there is a continuous variation of $G_g(t)$ patterns making

it difficult to find distinct areas. These problems would apply to clustering the gene expression data too.

Another potential problem is the case of co-regulated genes. If it is assumed that a gene is co-regulated as a linear combination of its transcription factors, then $G_g(t)$ will be a linear combination of two distinct activity profiles. As the proportion of effect that each transcription factor has on the co-regulated genes will vary from gene to gene this means that in this situation it is unlikely that well formed clusters will be found.

8.6 Detection of the principal activities of the response

8.6.1 Introduction

From section 8.5 it is clear that a more specialised approach is needed to gain additional information from the $G_g(t)$ data set. In particular it would be useful to be able to find out the number and shape of the main transcriptional activity profiles at work in a particular response. Additionally, it would be interesting to be able to find potential training sets for each of these transcriptional activity profiles so that the methods described in section 8.4 could be used without prior knowledge of the transcription factors or targets.

A potential way to meet these objectives would be to find small, tight, highly correlated groups of genes within the $G_g(t)$ data set. This would, in effect, be clustering but each gene would have to pass a high correlation threshold before it is considered a member of a cluster. The vast majority of genes will not be assigned to any cluster, which means that the problem of the large pool of genes that link the clusters and generally make the clusters unclear is removed. The tight clusters found can be associated with the principal transcription activity profiles that are active in the response. For this association to be valid principal transcription activity profiles have to affect a large number of genes and produce a strong response. These conditions are necessary to overcome the noise in the data and guarantee that a small tight cluster will be found. Both these conditions are biologically reasonable as the mechanisms that have the greatest effect on the response will meet these conditions, for example p53 affects a large number of genes and produces a strong response after DNA damage.

Each of the clusters found can be considered a training set for a transcription activity profile. The clusters will not pick up every target but will be good enough to produce a training set that can be used to find other likely targets of that transcription activity profile. The training sets can also be used to find the shapes of the principal driving forces of the response and may allow the identification of the genes behind this response. Also, the number of clusters found gives an idea of how many different subsystems are at work in the response. All this information could be helpful in reconstructing the network of the response.

8.6.2 Implementation

The problem of finding tight highly correlated groups of genes is best described in the terms of graph theory (Hartsfield and Ringel, 1990). A unidirected graph $G = (V, E)$ is made up of V, a set of vertices or nodes and E, a set of edges of lines that join two nodes. In this case the nodes are genes (or more specifically Affymetrix probe sets) and the edges indicate that the two genes have $G_g(t)$s that are highly correlated. An edge is formed between gene i and j if,

$$\langle G_i(t), G_j(t) \rangle > \alpha,$$

where α is a threshold that is set high. This produces a graph of genes with connections that indicate a shared activity profile driving their mRNA level. The whole graph will be split into groups that are not connected to each other and there will be many genes that are not connected at all. These groups could be considered as clusters that share the same transcription activity profile but a harsher criterion needs to be applied to get a robust gene group.

To achieve the required stringency, it was determined that $G_g(t)$ of a gene should be internally correlated. This means that for a gene to be included in the graph, the $G_g(t)$ for each of its replicates must be correlated to all other replicates to a value above a threshold, β. For example, if there are three replicates and $G_{g,r}(t)$ is the transcription activity profile for replicate r then,

$$\begin{aligned} \langle G_{g,1}(t), G_{g,2}(t) \rangle &> \beta, \\ \langle G_{g,1}(t), G_{g,3}(t) \rangle &> \beta, \\ \langle G_{g,2}(t), G_{g,3}(t) \rangle &> \beta, \end{aligned}$$

must hold for gene g to be included in the graph. A principal transcription activity profile will have a similar shape for each biological replicate. This criterion is not strict and β will generally be considerably lower than α threshold, but it effectively removes a considerably amount of the noisy $G_g(t)$s that can affect the results. In particular low level $G_g(t)$s who through correlation have their noise amplified are removed.

A clique within a graph is a collection of nodes V, such that for each two nodes there is an edge between them. This is equivalent to saying that V forms a subgraph that is a *complete graph.* In the context used here a clique is a group of genes where each gene has a $G_g(t)$ that is highly correlated with *every* other gene in the group. A clique in the constructed graph means that a particular transcription activity profile is appearing with a high reliability and that there is a group of genes that are very close to that profile and each other. These cliques are used as the basis for the formation of the clusters. If cliques were not used it is possible that a group of connected genes could have two genes on the outside of the group that had $G_g(t)$s that were very poorly correlated to each

other i.e. if the group was a long thin shape. A maximum clique is a clique that is not contained within any other clique in the graph and are the cliques that are of interest here. Finding maximum cliques in a graph is known as the maximum clique problem. The algorithm used to find the maximum cliques in the constructed graph is the effective Bron-Kerbosch algorithm (Bron and Kerbosch, 1973)[7].

If these maximum cliques share one or more nodes then the cliques are merged. This is to avoid transcription activity profiles that are very similar from being treated separately. If the number of nodes in the merged clique is four or greater then the clique is considered a cluster that contains a principal transcription activity profile.

8.6.3 Results and analysis of the transcription activity shapes

The procedure described in section 8.6.2 was performed on the $G_g(t)$s found from the 5Gy microarray data using only genes that appear in the clustering gene list (section 8.4.2). α was set to 0.85 and β was set at 0.5. Figure 8.17 displays the computed graph. There are five merged cliques (Figure 8.18) implying that there are five principal effects at work in the DNA damage response within the confines of the transcription model. The members of each merged clique will be examined below.

Merged clique 1

The first merged clique consists of 39 genes (see Table 8.8). This is considerably more than any of the other cliques indicating that the activity profile affects many genes and that the shape is very distinctive. The representative profile is produced by averaging the rescaled $G_g(t)$s (Figure 8.19). The shape is indicative of a strong early response, the activity rapidly rises and peaks at 2 hours before returning to its approximate initial value at 4 hours. The resolution of the plot is 2 hours, so it is possible that the actual peak could be between 0 and 2 hours. The most probable explanation for this activity is that it represents the activity of the AP-1 transcription factor. AP-1 is known to have a strong early response to DNA damage (Hayakawa *et al.*, 2004) and many of the member genes are known to be AP-1 targets including TNFAIP3 (Hayakawa *et al.*, 2004), TNFSF10 (Zou *et al.*, 2004), CD83 (Kim *et al.*, 2004), JUND (Yazgan and Pfarr, 2002) and JUN (Yazgan and Pfarr, 2002). AP-1 is an important transcription factor that is involved in cellular proliferation, transformation and death (Shaulian and Karin, 2002). AP-1 is not a single protein, but a dimer that can be formed from basic region-leucine zipper proteins that are members of the Jun, Fos, Maf and ATF sub-families (Shaulian and Karin, 2002). The most potent transcriptional activator in this group is c-Jun (Ryseck and Bravo, 1991). This clique might also contain other transcription target profiles that

[7]The Bron-Kerbosch algorithm computes all cliques of a graph using a branch-and-bound technique. It is efficient because it cuts off branches of the search tree that will not lead to new cliques at a very early stage. This is done by extending a trial clique in such a way that the bound condition (that the clique cannot be extended) becomes true at the earliest possible stage.

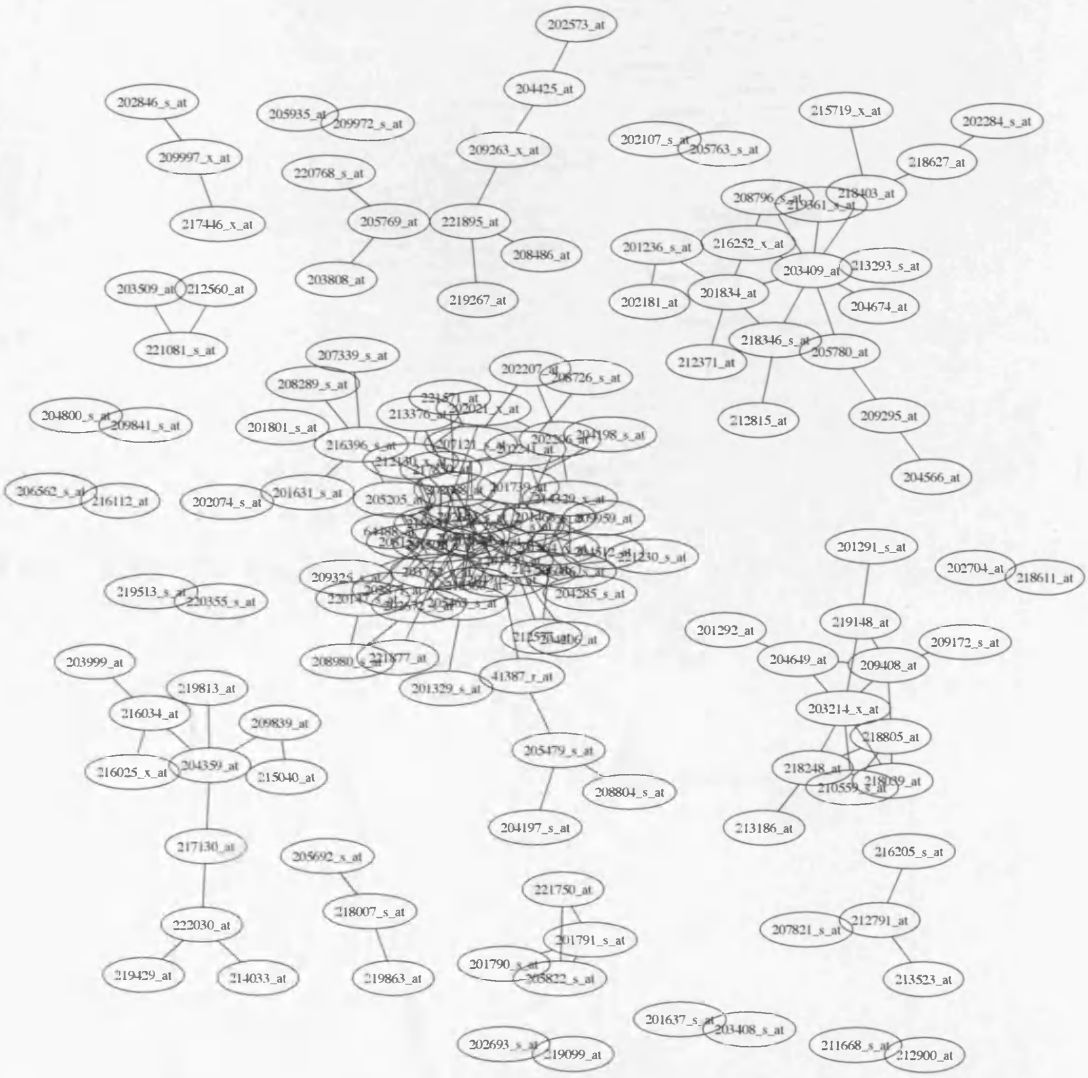


Figure 8.17: A graph produced from the DNA damage response $G_g(t)$ data where the nodes are a internally correlated subset from 2355 genes with $\beta > 0.5$ and the edges represent a correlation between $G_g(t)$s greater than 0.85.

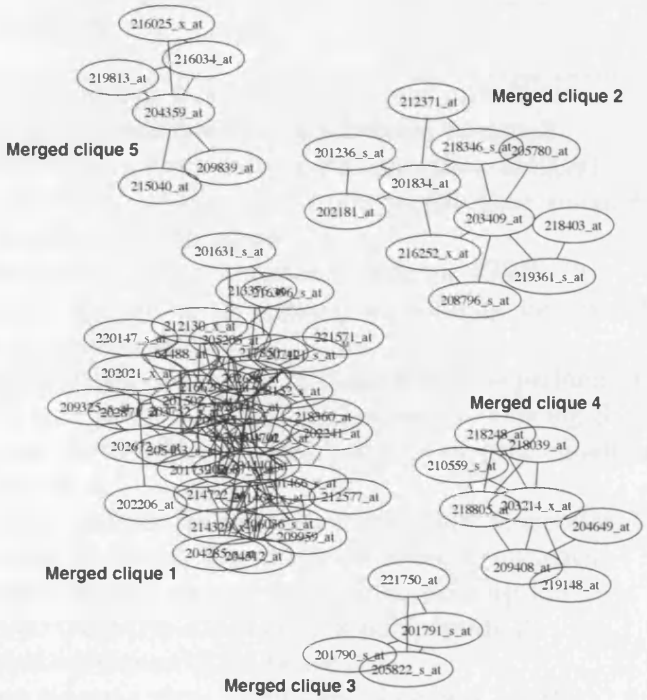


Figure 8.18: The graph of the final merged cliques found in the 5Gy $G_g(t)$ data from a subset of 2355 genes.

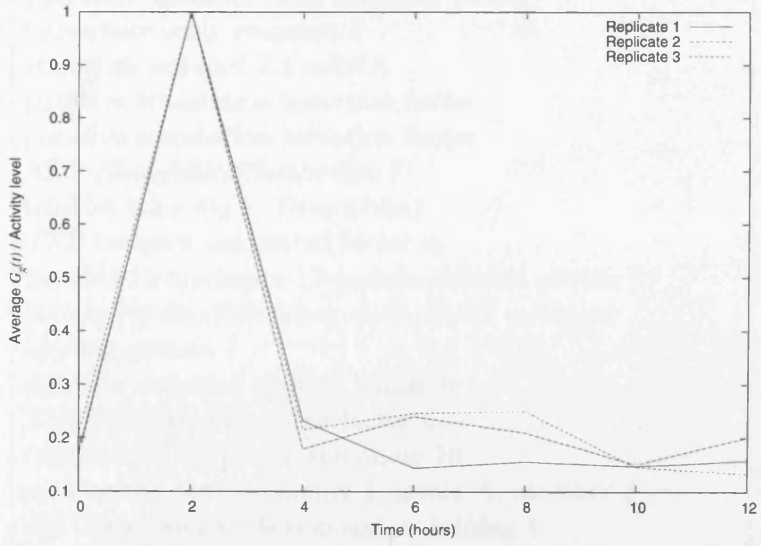


Figure 8.19: The average (after rescaling) of $G_g(t)$ for genes in merged clique 1. This represents the transcription activity profile for this set.

Table 8.8: The members of merged clique 1

Affymetrix tag	Description	Gene Symbol
202643_s_at	tumour necrosis factor, alpha-induced protein 3	TNFAIP3
202644_s_at	tumour necrosis factor, alpha-induced protein 3	TNFAIP3
209795_at	CD69 antigen (p60, early T-cell activation antigen)	CD69
201502_s_at	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha	NFKBIA
216620_s_at	Rho guanine nucleotide exchange factor (GEF) 10	ARHGEF10
202688_at	tumour necrosis factor (ligand) superfamily, member 10	TNFSF10
203752_s_at	jun D proto-oncogene	JUND
204440_at	CD83 antigen (activated immunoglobulin superfamily)	CD83
205205_at	v-rel reticuloendotheliosis viral oncogene homolog B, nuclear factor of kappa light polypeptide gene enhancer in B-cells 3	RELB
202687_s_at	tumour necrosis factor (ligand) superfamily, member 10	TNFSF10
217850_at	guanine nucleotide binding protein-like 3 (nucleolar)	GNL3
205463_s_at	platelet-derived growth factor alpha polypeptide	PDGFA
208152_s_at	DEAD (Asp-Glu-Ala-Asp) box polypeptide 21	DDX21
202672_s_at	activating transcription factor 3	ATF3
201464_x_at	v-jun sarcoma virus 17 oncogene homolog (avian)	JUN
214722_at	Notch homolog 2 (Drosophila) N-terminal like	NOTCH2NL
204702_s_at	nuclear factor (erythroid-derived 2)-like 3	NFE2L3
201466_s_at	v-jun sarcoma virus 17 oncogene homolog (avian)	JUN
206036_s_at	v-rel reticuloendotheliosis viral oncogene homolog (avian)	REL
214329_x_at	tumour necrosis factor (ligand) superfamily, member 10	TNFSF10
201739_at	serum/glucocorticoid regulated kinase	SGK
218360_at	RAB22A, member RAS oncogene family	RAB22A
201631_s_at	immediate early response 3	IER3
216396_s_at	etoposide induced 2.4 mRNA	EI24
202021_x_at	putative translation initiation factor	SUI1
212130_x_at	putative translation initiation factor	SUI1
202206_at	ADP-ribosylation factor-like 7	ARL7
202241_at	tribbles homolog 1 (Drosophila)	TRIB1
202871_at	TNF receptor-associated factor 4	TRAF4
204285_s_at	phorbol-12-myristate-13-acetate-induced protein 1	PMAIP1
204512_at	human immunodeficiency virus type I enhancer binding protein 1	HIVEP1
207121_s_at	mitogen-activated protein kinase 6	MAPK6
221571_at	TNF receptor-associated factor 3	TRAF3
209325_s_at	regulator of G-protein signalling 16	RGS16
209959_at	nuclear receptor subfamily 4, group A, member 3	NR4A3
213376_at	zinc finger and BTB domain containing 1	ZBTB1
212577_at	structural maintenance of chromosomes flexible hinge domain containing 1	SMCHD1
220147_s_at	family with sequence similarity 60, member A	FAM60A
64488_at	CDNA FLJ38849 fis, clone MESAN2008936	—

share similar transcription activity profiles, due to the two hour resolution, it is likely that different early responses would not be separated. For example there are a number of genes in the clique that appear to be involved with the NFκB pathway; RELB is a possible constituent part of the NFκB (Bours *et al.*, 1994) complex and NFKBIA (Sigala *et al.*, 2004) and IER3 (Pietzsch *et al.*, 1998) are both targets.

Merged clique 2

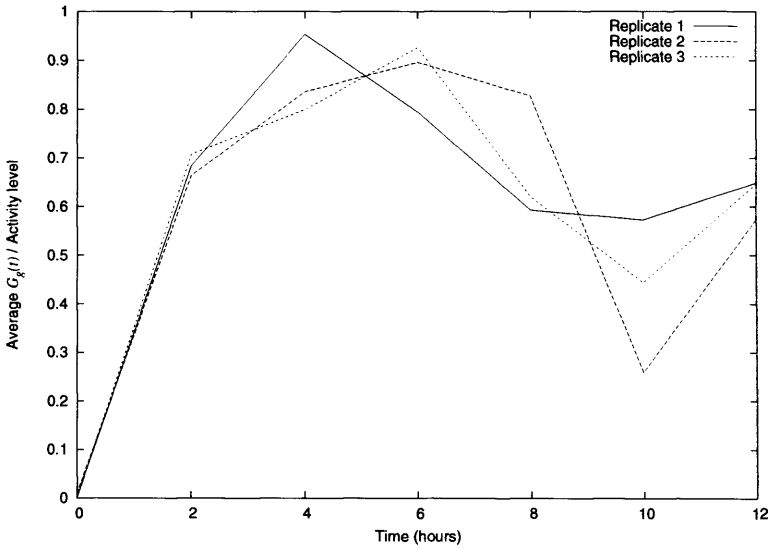


Figure 8.20: The average (after rescaling) of $G_g(t)$ for genes in merged clique 2. This represents the transcription activity profile for this training set.

Table 8.9: The members of merged clique 2

Affymetrix tag	Description	Gene Symbol	Verification Score
201236_s_at	BTG family, member 2	BTG2	2.50
201834_at	protein kinase, AMP-activated, beta 1 non-catalytic subunit	PRKAB1	6.30
202181_at	KIAA0247	KIAA0247	2.22
203409_at	damage-specific DNA binding protein 2	DDB2	10.7
216252_x_at	Fas (TNF receptor superfamily, member 6)	FAS	4.54
218346_s_at	sestrin 1t	PA26	3.90
212371_at	CGI-146 proteint	PNAS-4	2.61
205780_at	BCL2-interacting killer	BIK	6.57
208796_s_at	cyclin G1	CCNG1	5.18
218403_at	p53-inducible cell-survival factor	P53CSV	7.75
219361_s_at	hypothetical protein FLJ12484	FLJ12484	5.43

The shape of the representative $G_g(t)$ profile (Figure 8.20) and the members (Table 8.9) of merged clique 2 suggests that this clique represents p53 activity. Three genes,

DDB2, PA26 and BIK out of the five used in the original training set in section 8.4 are detected in this clique. This indicates that the automated procedure to find training set genes is working well. Out of the rest of the genes cyclin G1 (Bates *et al.*, 1996), P53CSV (Park and Nakamura, 2005), FAS (Li *et al.*, 2004a) and BTG2 (Rouault *et al.*, 1996) are all confirmed p53 targets. This means that the detection routine has picked well known p53 targets for seven out of the eleven training set genes. Also *all* of the genes have high p53 target verification scores. This is strong confirmation that the detection routine is working well and that is possible for the major activity profiles of the response to be detected. It is interesting that BTG2 appears as it was at a quite low rank in the predicted p53 target list found in section 8.4 at rank 62, this could indicate that the correlation method is quite sensitive to its training set.

Merged clique 3

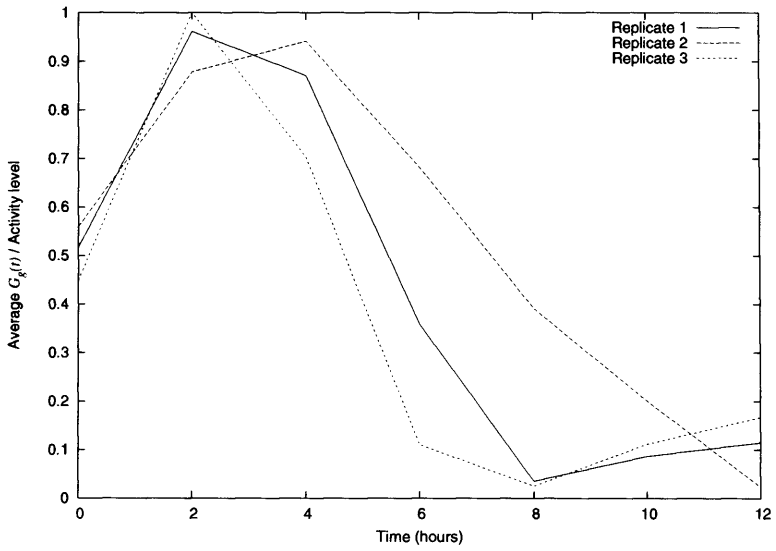


Figure 8.21: The average (after rescaling) of $G_g(t)$ for genes in merged cluster 3. This represents the transcription activity profile for this set.

Table 8.10: The members of merged clique 3

Affymetrix tag	Description	Gene Symbol
201790_s_at	7-dehydrocholesterol reductase	DHCR7
201791_s_at	7-dehydrocholesterol reductase	DHCR7
205822_s_at	3-hydroxy-3-methylglutaryl-Coenzyme A synthase 1 (soluble)	HMGCS1
221750_at	3-hydroxy-3-methylglutaryl-Coenzyme A synthase 1 (soluble)	HMGCS1

Merged clique 3 is made up of 2 genes, each with two Affymetrix probe sets in the group (see Table 8.10). Neither DHCR7 or HMGCS1 is known to be associated with the DNA damage response. The shape of the representative profile (Figure 8.21) seems to be

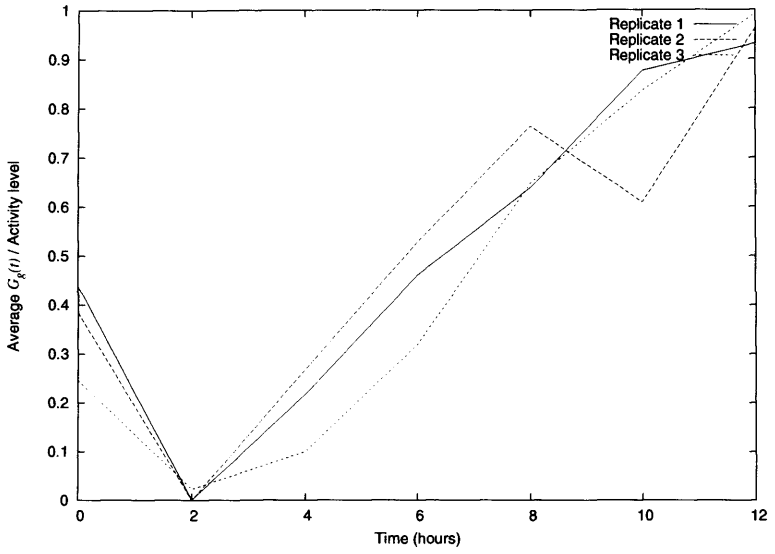


Figure 8.22: The average (after rescaling) of $G_g(t)$ for genes in merged clique 4. This represents the transcription activity profile for this set.

Table 8.11: The members of merged clique 4

Affymetrix tag	Description	Gene Symbol
203214_x_at	cell division cycle 2, G1 to S and G2 to M	CDC2
218805_at	GTPase, IMAP family member 5	GIMAP5
210559_s_at	cell division cycle 2, G1 to S and G2 to M	CDC2
218039_at	nucleolar and spindle associated protein 1	NUSAP1
218248_at	FLJ22794 protein	FLJ22794
209408_at	kinesin family member 2C	KIF2C
204649_at	trophinin associated protein (tastin)	TROAP
219148_at	PDZ binding kinase	PBK

in-between that of the AP-1/NFκB and p53 clique. It peaks between 2 and 4 hours, then rapidly declines and ends at a level significantly lower than its initial value. This merged clique could represent co-regulation by p53 and AP-1/NFκB but due to the low number of genes represented and hence lack of information it is difficult to judge the function. It is possible that his merged clique does not represent a principal transcriptional activity in the DNA damage response.

Merged clique 4

The fourth merged clique has a very distinctive profile (Figure 8.22), it rapidly diminishes to a trough at 2 hours and then increases at about the same rate moving past the starting value at about 5 hours. There are eight members of this training set (Table 8.11). CDC2 (also called CDK1) is a kinase that induces entry into mitosis (Lee *et al.*, 1988) and interestingly it is inhibited by p21, a p53 target (Yu *et al.*, 1998). This could explain why

Table 8.12: The members of merged clique 5

Affymetrix tag	Description	Gene Symbol
204359_at	fibronectin leucine rich transmembrane protein 2	FLRT2
216034_at	suppressor of hairy wing homolog 1 (Drosophila)	SUHW1
216025_x_at	cytochrome P450, subfamily IIC	–
219813_at	LATS, large tumour suppressor, homolog 1 (Drosophila)	LATS1
215040_at	Hypothetical protein FLJ11712	FLJ11712
209839_at	dynamin 3	DNM3

there is an initial drop if one of its *negative* regulators is rising but it does not explain the subsequent rise. The gene expression model only deals with positive regulation so care must be taken. Two other genes NUSAP1 and KIF2C (Kim *et al.*, 1997) are associated with the mitotic spindle and moving cargo along microtubules respectively. This suggests that the activity behind this training set could be associated with the G2/M cell cycle phase or mitosis. The cell cycle is arrested in response to DNA damage (Vousden, 2000) so this could explain the initial drop, the increase above the initial value could be a more subtle effect. All the gene expression levels measured are based on an average of the population and the levels are normalised so that the total expression level of a microarray is the same. Initially, there will be a mixed population of cells on the path to apoptosis or survival, so the gene expression signal will be a mixture of the two. As cells commit apoptosis and disappear from the population, a greater proportion of the signal will be represented by the surviving genes and so cell cycle associated mRNA levels will go up too. So the $G_g(t)$ profiles could represent simply the number in G2/M phase. An alternative could be that DNA damage triggers a checkpoint that arrests the cells and lasts for two hours, as the checkpoint is released all the cells are piling up in G2/M phase and so an increased response is seen. These two possibilities could be tested by microarray time series experiments on cell mutants that cannot arrest. Interestingly, active p53 peaks after mitosis appears to resume. Early cell-cycle arrest is independent of p53, but p53 can prolong the arrest (Wahl *et al.*, 1997). So in this case it appears that p53-dependent arrest is not occurring, suggesting that once a cell is in the process of apoptosis mitosis is not prevented.

Merged clique 5

The final training set has a distinctive two peak shape that peaks at 2 and 8 hours (see Figure 8.23). It is made up of 6 genes (see Table 8.12). The transcription activity this training set represents is unclear as there is not much information about its members. LATS1 is both a tumour suppressor and a cell-cycle regulator binding to CDC2 briefly in mitotic cells (Kemp, 1999). It plays a role in the exit from mitosis.

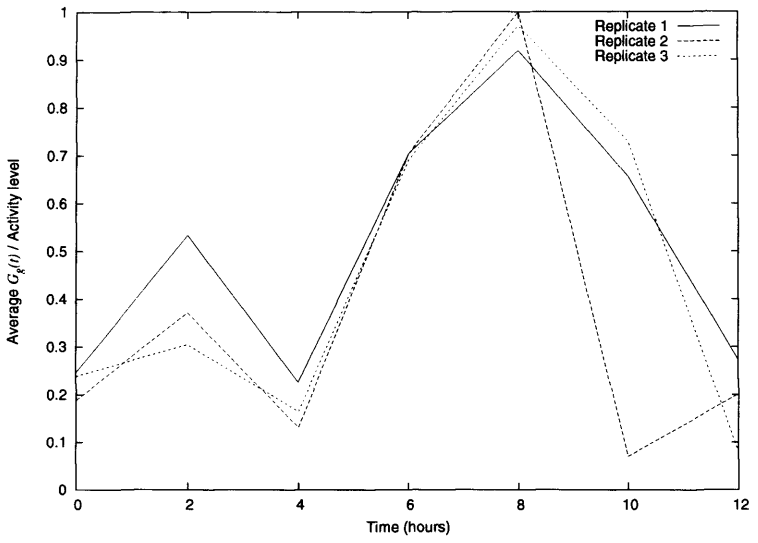


Figure 8.23: The average (after rescaling) of $G_g(t)$ for genes in training set 5.

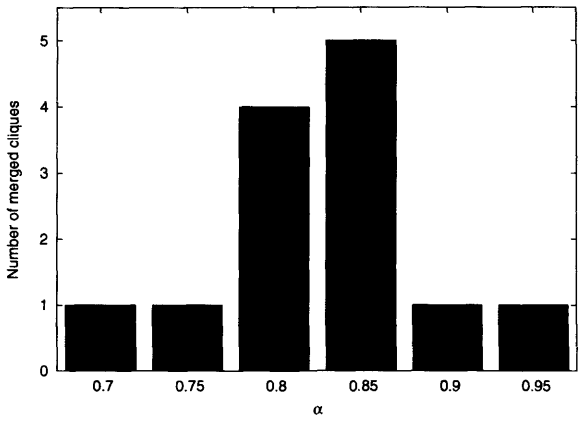
Robustness of the merged cliques

There are a number of thresholds that are set in this procedure and it is important that the results are robust to alterations in these parameters. To test the robustness training sets were found with α (the connection threshold) set at a range of different values between 0.7 and 0.95, β remained equal to 0.5. The number of cliques varies with α but only at $\alpha = 0.8$ or 0.85 are there more than one clique (Figure 8.24(a)). If α is set too low the genes will all have connections between them and so one large network is formed, but if α is set too high, then only a few genes are linked. The average number of genes per training set decreases as α is increased (Figure 8.24(b)). When choosing α one has to ensure that there is enough information in the network, but not so much that noise affects the separation. This suggests that a good policy to take would be to choose α so that it maximises the number of training sets. The training sets formed when $\alpha = 0.8$ rather than $\alpha = 0.85$ are very similar except that the third merged clique has combined suggesting that these merged cliques have a closely related function. Interestingly, merged clique two, which represents p53 activity has all five training genes used in section 8.4 as members when $\alpha = 0.8$.

The results when β (the internal correlation threshold) was varied were examined with α kept constant at 0.85. When $\beta = 0.5$ and $\beta = 0.25$ there were five merged cliques and when $\beta = 0$ there were six. The members of the five common merged cliques were surprisingly robust apart from clique 5 (see Table 8.13).

Figure 8.25 shows the representative transcription activity profile for the new merged clique created at $\beta = 0$. It has six member genes. The individual replicates do not have very similar profiles and the overall behaviour is the same as for merged clique 4. It seems likely that this is an aberration and should not be considered a principal activity profile.

(a)



(b)

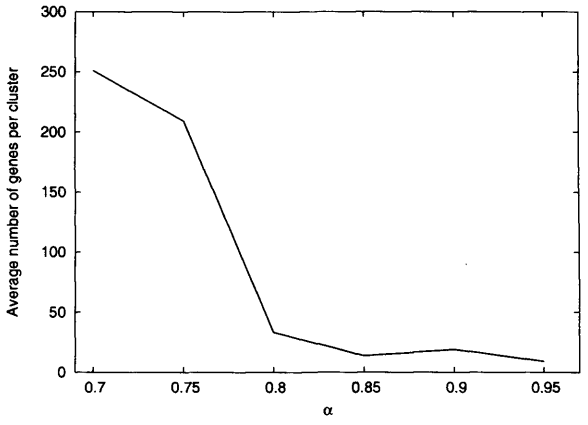


Figure 8.24: Plots showing how (a) the number of merged cliques and (b) the average number of genes per clique vary with changes in α, the graph connection threshold.

Table 8.13: A table showing how the numbers of members of the merged cliques vary with β, the internal correlation threshold.

Merged clique no.	$\beta = 0$	$\beta = 0.25$	$\beta = 0.5$
1	43	43	39
2	11	11	11
3	4	4	4
4	8	8	8
5	227	71	6

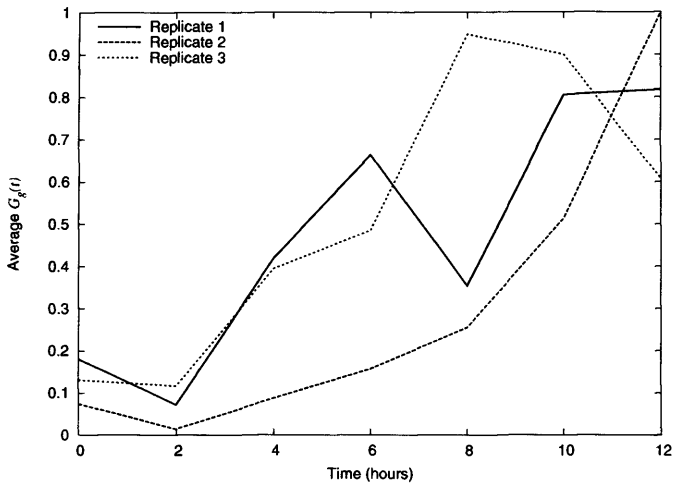


Figure 8.25: The representative profile of a new training set found when $\beta = 0$ and $\alpha = 0.85$.

Overall, the method to find training sets from $G_g(t)$ is extremely robust to parameter change.

8.6.4 Predicting target genes using each training set

The members of the five merged cliques can be used as training sets in the method described in section 8.4.3, to identify gene targets that are controlled by the same transcription activity. This process will produce a larger number of genes associated with the activity profile helping to identify the source of the activity.

Appendix D.3 displays the ranked lists for each training set for genes that have a correlation of at least 0.7 with the representative profile. Training set two is associated with p53 activity and has been fairly well analysed already. The accuracy of the results can be judged by comparing with the verification scores. It works well at first with 80% of the top 50 being confirmed as p53 targets (verification score > 1). Only 58% are confirmed p53 targets in the top 100 though, which is significantly worse that the 71% of the top 100 confirmed in section 8.4. It is unclear why this occurs as the representative profiles in both cases are fairly similar (See Figure 8.10(a) & 8.20) and three genes are shared between the training sets. It does show that the correlation procedure's performance can be sensitive to the members of the training set.

For training set three there are only 27 genes that correlate with the representative profile with a correlation greater that 0.7. Also 10 of these genes appear in the AP-1 list which suggests that this activity is part of the AP-1 activity rather than an independent activity. This along with the small number of genes in the training set and the fact it merges with the training set 1 when the threshold α is decreased suggests that training set 3 does not represent a principal activity profile.

To identify common functionality in the clusters a website based program called

Table 8.14: Ranked list of the likely functions of genes found to be transcribed by activation profile 1 (p-value < 0.1).

GO Biological Processes Definition	Count	Percentage	p-value
REGULATION OF BIOLOGICAL PROCESS	27	38.6	9.67×10^{-7}
APOPTOSIS	11	15.7	2.44×10^{-6}
REGULATION OF CELLULAR PROCESS	12	17.1	9.12×10^{-5}
REGULATION OF APOPTOSIS	7	10	0.000255
REGULATION OF PHYSIOLOGICAL PROCESS	18	25.7	0.000923
REGULATION OF TRANSCRIPTION, DNA-DEPENDENT	16	22.9	0.00174
REGULATION OF METABOLISM	17	24.3	0.00177
REGULATION OF SIGNAL TRANSDUCTION	5	7.1	0.00262
TRANSCRIPTION FROM POL II PROMOTER	7	10	0.00484
ANTI-APOPTOSIS	4	5.7	0.00506
CELLULAR PHYSIOLOGICAL PROCESS	26	37.1	0.00506
I-KAPPAB KINASE/NF-KAPPAB CASCADE	4	5.7	0.00693
CELLULAR PROCESS	35	50	0.00721
NUCLEOBASE, NUCLEOSIDE, NUCLEOTIDE AND NUCLEIC ACID METABOLISM	20	28.6	0.00756
METABOLISM	37	52.9	0.00855
RESPONSE TO STRESS	9	12.9	0.0100
POSITIVE REGULATION OF I-KAPPAB KINASE/NF-KAPPAB CASCADE	3	4.3	0.0323
SIGNAL TRANSDUCTION	17	24.3	0.0428
RESPONSE TO PATHOGENIC BACTERIA	2	2.9	0.0542
PROTEIN KINASE CASCADE	4	5.7	0.0577
INDUCTION OF APOPTOSIS	3	4.3	0.0721
CELL COMMUNICATION	19	27.1	0.0729
POSITIVE REGULATION OF APOPTOSIS	3	4.3	0.0781
RESPONSE TO PATHOGEN	2	2.9	0.0835

DAVID (Dennis *et al.*, 2003) was applied to the ranked lists. This program assigns genes to functions using the Gene Ontology biological processes definitions and then ranks them according to the EASE-score, a method that identifies functional categories over-represented in a gene list relative to the representation within the proteome of a given species. This should give a good idea of the functions that are affected by the principal transcription activities. The Gene Ontology definitions are hierarchical so in the results some functional definitions are contained with other parent definitions. The training set 1 genes suggested that the AP-1 transcription factor was regulating the activity of this group. Many of the top identified functions linked with the ranked list based on the transcriptional activity of training target set 1 (Table 8.14) support the association with AP-1; with functions such as the "response to stress", "signal transduction", "apoptosis" and "anti-apoptosis". Interestingly, there is also a couple of functions also associated with the NFκB pathway. p53 was associated with training set 2 and the corresponding function list supports that association p53 (Table 8.15). All the main functions of p53 are covered including induction of apoptosis, DNA repair and cell cycle arrest.

The top functions for the targets of training set 4 provide evidence that this transcription activity profile is associated with mitosis (see Table 8.16). The majority of the functions are linked with mitosis with functions such as "mitosis", "chromosome condensation" and "regulation of the cell cycle". This shows the power of this approach and provides further evidence that the principal transcriptional activities are identifying behaviours of interest. It is less clear what the top functions for training set 5 show (see Table 8.17) but the majority of the functions are associated with cell signalling and in particular regulation of these processes. Especially important seems to be inter-cell communication and cell adhesion. Interestingly, there are some NFκB functions in the list. Maybe this transcriptional activity is associated with the recovery of the cell and

Table 8.15: Ranked list of the likely functions of genes found to be transcribed by activation profile 2 (p-value < 0.1).

GO Biological Processes Definition	Count	Percentage	p-value
REGULATION OF CELLULAR PROCESS	11	27.5	1.70×10^{-6}
INDUCTION OF APOPTOSIS	6	15	4.12×10^{-6}
RESPONSE TO DNA DAMAGE STIMULUS	7	17.5	4.96×10^{-6}
CELL DEATH	8	20	2.52×10^{-5}
DNA REPAIR	6	15	3.81×10^{-5}
REGULATION OF APOPTOSIS	6	15	0.000123
NEGATIVE REGULATION OF CELL PROLIFERATION	5	12.5	0.000127
REGULATION OF CELL PROLIFERATION	6	15	0.000146
APOPTOSIS	7	17.5	0.000174
REGULATION OF BIOLOGICAL PROCESS	15	37.5	0.000506
RESPONSE TO STRESS	8	20	0.00114
REGULATION OF CYCLIN DEPENDENT PROTEIN KINASE ACTIVITY	3	7.5	0.00245
CELLULAR PHYSIOLOGICAL PROCESS	17	42.5	0.00442
PROTEIN MODIFICATION	9	22.5	0.00503
G1/S TRANSITION OF MITOTIC CELL CYCLE	3	7.5	0.00506
CELL PROLIFERATION	8	20	0.00548
CELL CYCLE ARREST	3	7.5	0.00580
DNA METABOLISM	6	15	0.00644
REGULATION OF CELL CYCLE	5	12.5	0.00756
MACROMOLECULE METABOLISM	14	35	0.00811
REGULATION OF ENZYME ACTIVITY	3	7.5	0.0134
PROTEIN METABOLISM	12	30	0.0137
METABOLISM	22	55	0.0203
NUCLEOTIDE-EXCISION REPAIR	2	5	0.0493
MITOTIC CELL CYCLE	3	7.5	0.0576
INDUCTION OF APOPTOSIS BY EXTRACELLULAR SIGNALS	2	5	0.0585
PHOSPHATE METABOLISM	6	15	0.0591
CELL CYCLE	5	12.5	0.0670
APOPTOTIC PROGRAM	2	5	0.0802
RESPONSE TO STIMULUS	11	27.5	0.0816
CELLULAR PROCESS	19	47.5	0.0838

Table 8.16: Ranked list of the likely functions of genes found to be transcribed by activation profile 4 (p-value < 0.1).

GO Biological Processes Definition	Count	Percentage	p-value
CELL CYCLE	12	33.3	3.59×10^{-8}
MITOSIS	7	19.4	8.04×10^{-8}
CELL PROLIFERATION	12	33.3	1.42×10^{-6}
CHROMOSOME CONDENSATION	3	8.3	0.000389
REGULATION OF CELL CYCLE	5	13.9	0.00507
CELLULAR PROCESS	20	55.6	0.00748
CELL GROWTH AND/OR MAINTENANCE	14	38.9	0.00815
CELLULAR PHYSIOLOGICAL PROCESS	15	41.7	0.0101
MITOTIC CHROMOSOME CONDENSATION	2	5.6	0.0138
MITOTIC PROPHASE	2	5.6	0.0138
MITOTIC ANAPHASE	2	5.6	0.0224
PROTEIN AMINO ACID PHOSPHORYLATION	5	13.9	0.0253
PROTEIN MODIFICATION	7	19.4	0.0343
PROTEIN METABOLISM	10	27.8	0.0448
DNA REPLICATION	3	8.3	0.0551
REGULATION OF MITOSIS	2	5.6	0.0608
RESPONSE TO STRESS	5	13.9	0.0620
RESPONSE TO OXIDATIVE STRESS	2	5.6	0.0722
DNA METABOLISM	4	11.1	0.0925

signalling this recovery with other cells.

Finding the biological processes associated with each transcriptional activity profile has provided further insight into the usefulness of this approach. This methodology results in clusters of gene that share a distinct functional fingerprint and an associated activity profile. This methodology have extracted this information efficiently. No other approach manages to gain functional information like this.

8.6.5 Discussion of the detection of principal activities

A simple method of detecting small tight clusters of genes have been used to find genes with closely correlated $G_g(t)$s. This method is robust to parameter variation. With only

Table 8.17: Ranked list of the likely functions of genes found to be transcribed by activation profile 5 (p-value < 0.05).

GO Biological Processes Definition	Count	Percentage	p-value
CELLULAR PROCESS	234	42.4	1.45×10^{-5}
CELL COMMUNICATION	138	25	1.52×10^{-5}
POSITIVE REGULATION OF SIGNAL TRANSDUCTION	12	2.2	1.85×10^{-5}
INTRACELLULAR SIGNALING CASCADE	51	9.2	5.68×10^{-5}
POSITIVE REGULATION OF I-KAPPAB KINASE/NF-KAPPAB CASCADE	11	2	6.04×10^{-5}
REGULATION OF SIGNAL TRANSDUCTION	15	2.7	0.000159
I-KAPPAB KINASE/NF-KAPPAB CASCADE	12	2.2	0.000213
CELL ADHESION	35	6.3	0.000392
SIGNAL TRANSDUCTION	109	19.7	0.000433
CELLULAR PHYSIOLOGICAL PROCESS	153	27.7	0.000588
REGULATION OF CELLULAR PROCESS	39	7.1	0.000685
PROTEIN KINASE CASCADE	16	2.9	0.00360
PROTEIN METABOLISM	98	17.8	0.00602
PROTEIN TARGETING	10	1.8	0.00849
HETEROPHILIC CELL ADHESION	9	1.6	0.0107
PROTEIN TRANSPORT	21	3.8	0.0111
PROTEIN COMPLEX ASSEMBLY	9	1.6	0.0112
REGULATION OF BIOLOGICAL PROCESS	99	17.9	0.0120
CELL GROWTH AND/OR MAINTENANCE	127	23	0.0123
NEUROPEPTIDE SIGNALING PATHWAY	9	1.6	0.0124
CELL-CELL SIGNALING	25	4.5	0.0156
INTRACELLULAR PROTEIN TRANSPORT	15	2.7	0.0166
INTRACELLULAR TRANSPORT	20	3.6	0.0213
MACROMOLECULE METABOLISM	111	20.1	0.0259
CELL MOTILITY	17	3.1	0.0259
INTRACELLULAR RECEPTOR-MEDIATED SIGNALING PATHWAY	4	0.7	0.0286
NEUROGENESIS	19	3.4	0.0306
GLYCEROPHOSPHOLIPID METABOLISM	4	0.7	0.0317
CELL-CELL ADHESION	14	2.5	0.0363
DEVELOPMENT	62	11.2	0.0400
MEMBRANE LIPID METABOLISM	8	1.4	0.0400
RAS PROTEIN SIGNAL TRANSDUCTION	4	0.7	0.0421

a relatively small number of data points, a simple model of gene transcription and an estimation of one parameter per gene, it is possible to gain detailed information about a biological response using only mRNA data. This method can be used to detect training sets of genes without prior knowledge, that can be used in other methods. It also detects the number and profile shapes of the principal activities that are working at the protein level in the response. It is an improvement over the use of clustering because it ignores the noisy data that can make the partition of data error prone. Without the use of $G_g(t)$ data it would be difficult to have the global information from mRNA data.

Even though the cliques could just be used as training sets in either the correlation ranking method (section 8.4.3) or the method described by Barenco *et al.* (2005) to detect groups of transcription targets there are other possible uses. One interesting possibility is to decompose each gene's estimated $G_g(t)$ profile in terms of the representative profile of each training set. Each training set represents a principal activity in the response and so can be thought of as an axis in the profile space. In this way it would be possible to detect co-regulated genes and assign probabilities to a particular gene being associated with a particular activity. It is clear that the representative profiles would not be orthogonal and so a dual space would have to be created using a decomposition scheme such as Gram-Schmidt orthogonalisation (Heath, 1997). The Gram-Schmidt algorithm would produce different results depending on the order that the representative profiles are used, so the order would be chosen so that the representative profiles account for the maximum amount of all the genes $G_g(t)$s. Once the proportion of a gene's $G_g(t)$ has been assigned to each representative profile it would be simple to determine whether the gene belonged

to a particular activity or is co-regulated.

There are limitations to this approach as it is currently implemented. The major limitation is that here only genes that have a mRNA profile that is up-regulated are included in the analysis, this means that there could be positive regulation that is inhibited by the DNA damage response that is not being picked up. It is difficult to see ways to overcome this unless all genes that change are included, but this would introduce a large number of genes that are not described by the model and hence introduce additional noise. Another limitation is that it can only consider positively regulated genes. One problem is that it can be difficult to find the protein or proteins that are responsible for the activities that are found. In many cases this would need extra experiments and diligent searching of the literature. A final limitation is that this method only finds the principal activities behind the response, this is a requirement because there needs to be a high probability that a clique will be found and this will only happen if a large number of genes are affected in a significant way. The training sets found though are robust to parameter variation which is suggestive that this methods detects all the activities possible in the data.

This method enables the extraction of additional information from microarray data that would not be possible otherwise. It would be interesting to see this approach applied to a new system to see what activities would be observed. It would also be interesting to see whether increasing the number of data points would give the same number of activities with increased resolution, or whether new types of behaviour could be detected.

8.7 Principal components analysis of the G time profile

Principal components analysis (PCA) is a well established technique used in the analysis of multivariate data. PCA performs a optimal linear transformation on the data such that the largest amount of variance is along the first axis, the second largest amount of variance is along the second axis, and so forth (Wall *et al.*, 2002). The axes are orthonormal and the direction of the axes describes the principal components of the data (the direction of the first axis is the first principal component and so on). The main application of PCA is to allow the reduction in the number of dimensions of the data by eliminating the bottom principal components. This reduction still retains the greatest possible amount of variance in the data and thus the core characteristics of the data. This simplification can help in the visualisation of the data and detecting structure on the data. Additionally, the principal components themselves can provide information about the data.

The principal components are determined by finding the eigenvalues and eigenvectors of the covariance matrix. The eigenvectors are the principal components and the eigenvalues give the relative amount of variance explained by the associated eigenvector. Singular value decomposition (SVD) decomposes a $m \times n$ matrix A into three matrices

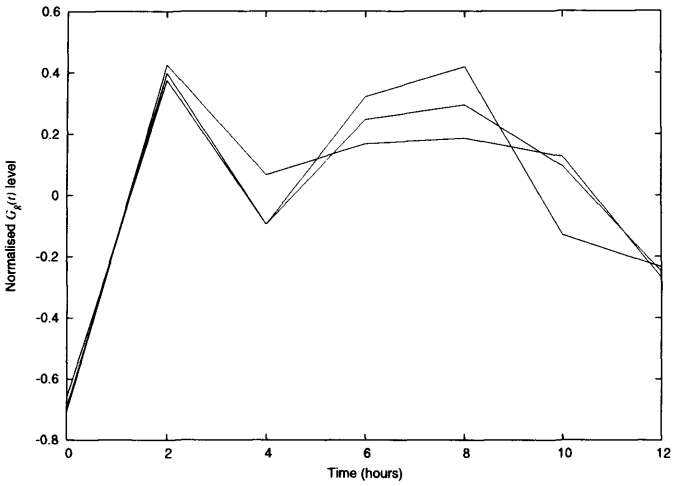


Figure 8.26: The mean $G_g(t)$ profile based on data from a 2355 subset of genes that have been filtered based on internal correlation and have been normalised to have a mean of zero and a variance of one.

(Press *et al.*, 2002),

$$A = USV^T,$$

where S is an $n \times n$ diagonal matrix containing the singular values, U is an $m \times n$ unitary matrix and V is an $n \times n$ unitary matrix. SVD is closely related to PCA and can be used to calculate the principal components if the data matrix is pre-processed by centering the columns. In this situation, the columns of V give the principal components and the singular values squared provide the associated proportion of variance described by each principal component. SVD/PCA is routinely applied to gene expression data, mainly to remove noise from the data, but it has recently been used to extract more information about pathways, detect patterns in gene expression and even reverse engineer networks (Wall *et al.*, 2001; Alter *et al.*, 2000; Holter *et al.*, 2000; Yeung *et al.*, 2002; Tomfohr *et al.*, 2005; Yeung and Ruzzo, 2001). In these situations it is normal to construct the matrix so that the genes are the rows and the columns are the experiments. Depending on the aim of the analysis either the experiments or the genes are treated as the variables of the PCA.

PCA was applied to the $G_g(t)$ profiles of the subset of 2355 genes. The idea is that principal activity profiles will cause a maximum amount of variance as the profile will be strong, distinct and present or absent. The data was filtered so that $G_g(t)$ profiles that do not have an internal correlation of 0.5 were removed (see section 8.6.2). Each of the remaining $G_g(t)$s were then normalised to have zero mean and a variance of one. The matrix was constructed with the columns representing the time points and the rows genes. Each column was centered by removing the column mean and SVD was applied to the resulting matrix. This removes the mean $G_g(t)$ profile (Figure 8.26).

The amount of variance explained by each principal component drops off rapidly

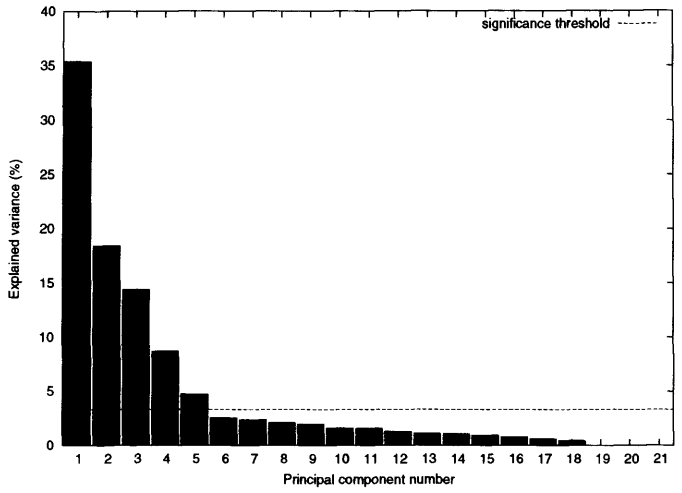


Figure 8.27: A plot showing how the percentage of data variance is contained along each principal component. This is based on PCA applied to the $G_g(t)$s of 2355 subset of genes that have been filtered based on internal correlation and have been normalised to have a mean of zero and a variance of one.

as the rank of the principal component increases (Figure 8.27). The later principal components are unlikely to provide useful information and so these are normally ignored focussing on the significant components. A heuristic approach is used to determine the significant components, first suggested by Everitt and Dunn (2001), which only accepts components as significant if a component explains a proportion of the variance of $0.7/n$ or greater, where n is the total number of components. In this case the first five components are significant (Figure 8.28). The replicates do not share the same shape for the sixth principal components and above, suggesting that a sensible threshold has been used. The principal components are axes and so each component multiplied by a positive *or* negative factor contribute to each $G_g(t)$ profile. Therefore when considering the shape of the principal components one must also consider the profile when reflected horizontally. For example, the sharp peak at 2 hours of the first principal component also represents a sharp trough at two hours.

Generally there seems to be no relationship between the principal components and the principal activities found in section 8.6.3 (Figures 8.28 and 8.29). The exception is the first principal component (Figure 8.28(a)) which could be assigned to principal activity 1, the activity profile associated with AP-1 (Figure 8.29). Interestingly, principal activity 1 had the most members in its corresponding clique. PCA and the method for finding cliques extract different information from the $G_g(t)$ space, PCA finds directions that account for the largest variations in the data whilst the latter finds areas in the space where the density of the points are very high. The principal components are also different from the principal activity profiles because the principal components are linearly independent while the principal activities are not. It is unclear whether the principal components in themselves are biologically meaningful and it has been shown in

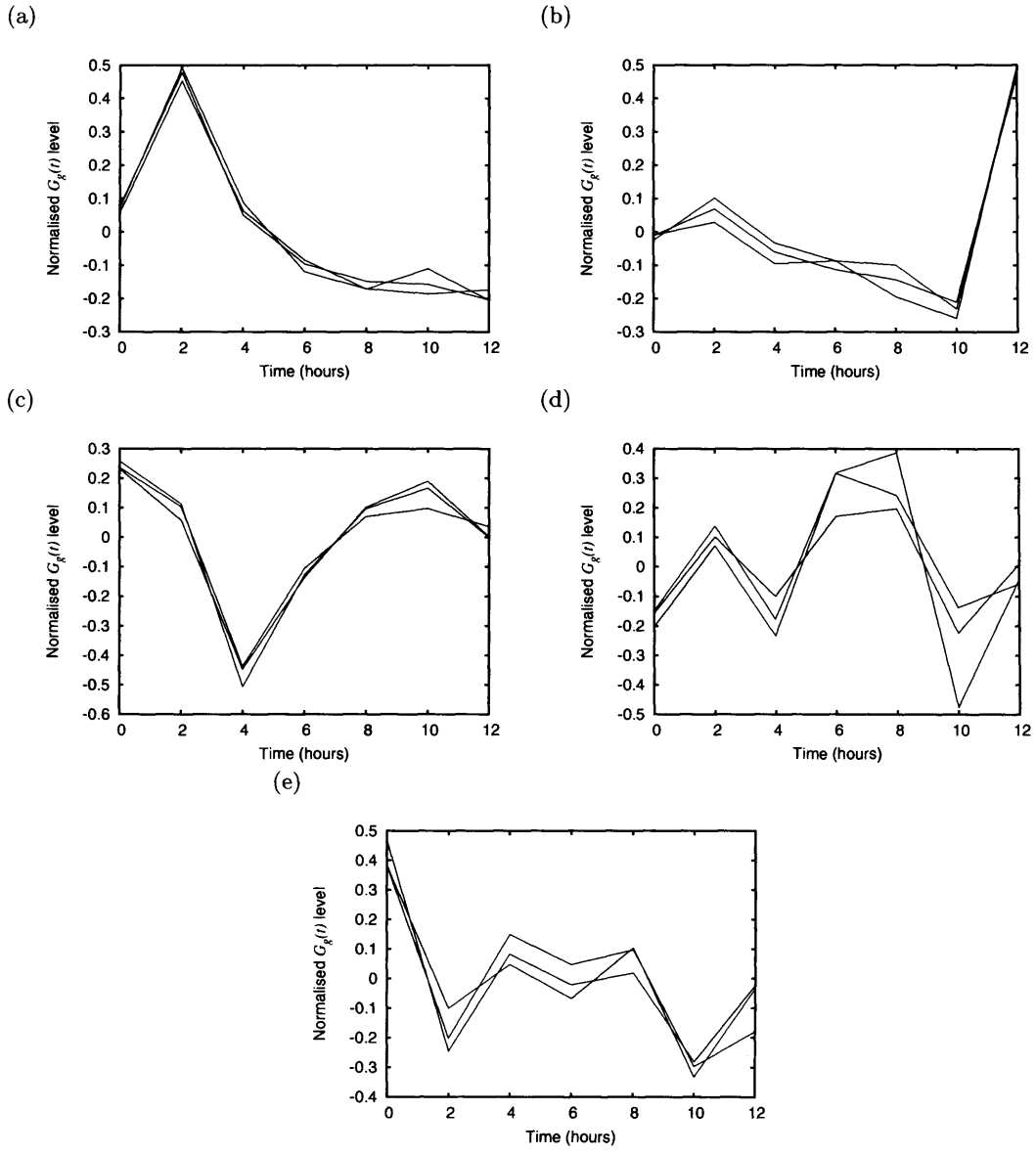


Figure 8.28: The principal components found from a subset of the $G_g(t)$ data of rank (a) 1, (b) 2, (c) 3, (d) 4 and (e) 5. The three lines represent the three replicates.

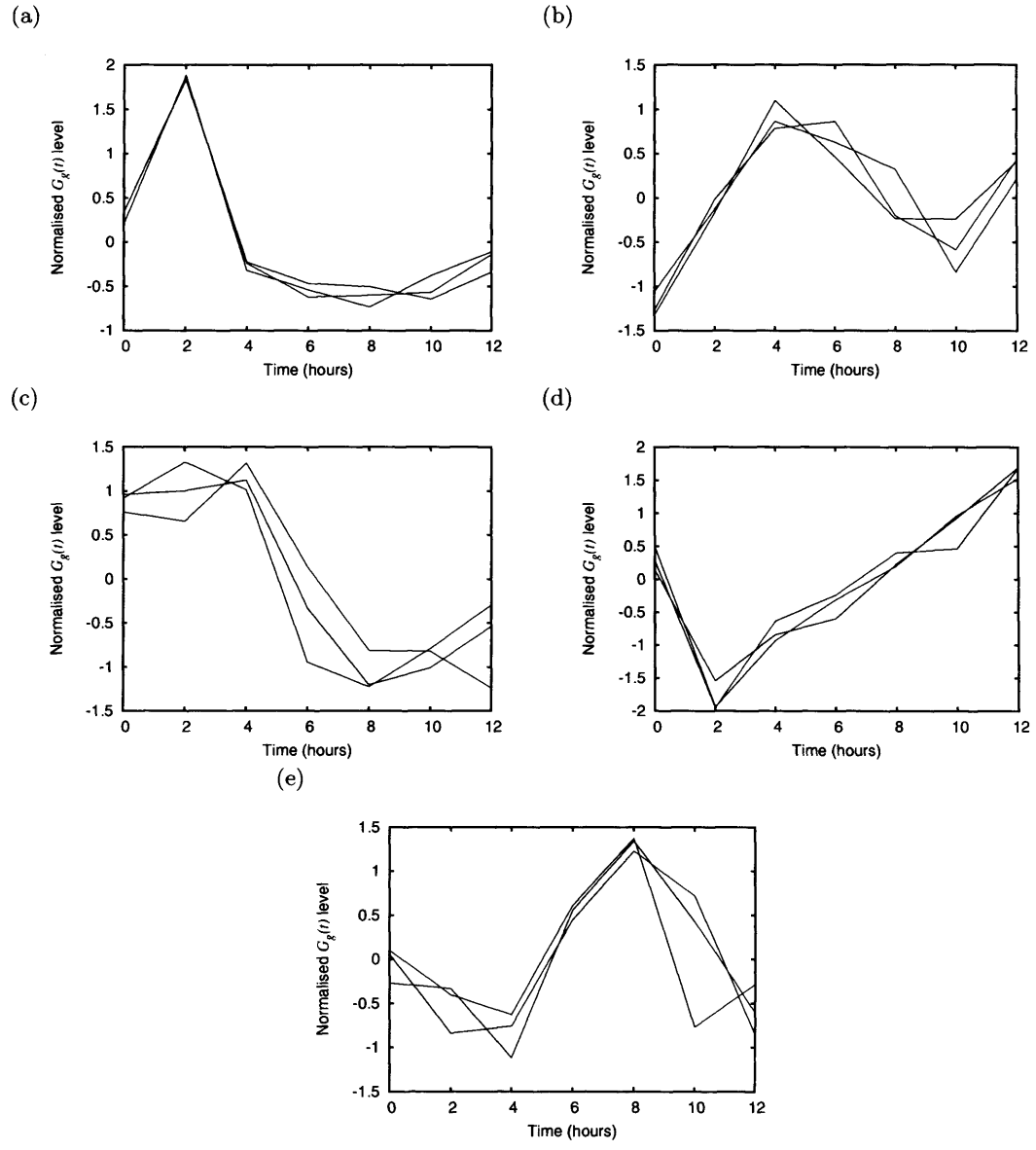


Figure 8.29: The principal activities (a) 1, (b) 2, (c) 3, (d) 4 and (e) 5 found in section 8.6.3. These have been adjusted to have a mean of zero and variance one and then had the average profile (Figure 8.26) removed. This allows comparison with the principal components. The three lines represent the three replicates.

Table 8.18: The co-ordinates of the five activity profiles found in section 8.6.3 in the space of principal components.

Principal activity profile	Principal components				
	1	2	3	4	5
1	3.11	0.878	0.848	-0.0652	-0.201
2	-0.377	0.569	-2.42	0.997	-0.624
3	3.17	0.148	-1.22	-1.15	1.77
4	-3.83	1.57	1.43	-1.13	0.258
5	-1.32	-1.40	1.13	1.74	0.434

simulated studies that the principal components are not always biologically meaningful (Wall *et al.*, 2002). Despite this the first principal component does seem to suggest that a sharp peak/trough is a distinctive feature of the response which agrees with the clique activity profiles with 4 out of 5 having a rapid change between 0 and 2 hours. This is biologically reasonable as one would expect a strong initial reaction to DNA damage.

The activity profiles have a distinctive profile in the principal component space (Table 8.18). In activity profile 1, 3 and 4 the first principal component (strong peak/trough at two hours) is dominant occurring in the positive direction for activity profile 1 and 3 and in the negative direction for the activity profile 4. Activity profile 2, the profile associated with p53, is dominated by the third component, which is a strong peak at 4 hours (in the negative direction). Activity profile 5, which has two peaks, requires a combination of the first four principal components in roughly equal portions showing that the activity profiles cannot be assigned one to one with a principal component. The activity profiles are the average of a number of genes belonging to the associated clique. Figure 8.30 shows how the cliques arrange themselves in the principal component space. Clique 3 is generally in the same area as clique 1 which may suggest that clique 3 is really part of clique 1. When principal component 1 and principal component 3 are used as axes clique 3 is slightly separated from clique 1 (Figure 8.30(b)). Interestingly IER3, which is known to be co-regulated by p53 (associated with clique 2) and NFκB (associated with clique 1) (Huang *et al.*, 2002; Im *et al.*, 2002), has co-ordinates (3.1,-1.6), which are very close to clique 3. This suggests that clique 3 may represent co-regulation by the activities behind clique 1 and 3 as was suspected from its activity profile (see section 8.6.3). The first principal component successful separates the cliques into three groups, but clique 2 and 5 hold similar values. The introduction of the second or third principal component separates these two. The first and third principal components separate the cliques to the greatest extent which suggests that the activity profiles can be described in terms of the peak/trough at two hours and the trough/peak at four hours.

The method of finding cliques and principal activity profiles is distinct from PCA, the former finds areas in the parameter space where the density of the points are very high whereas PCA finds directions that account for a large variation in the data. The activity profiles directly provide biological information whereas there is no guarantee that

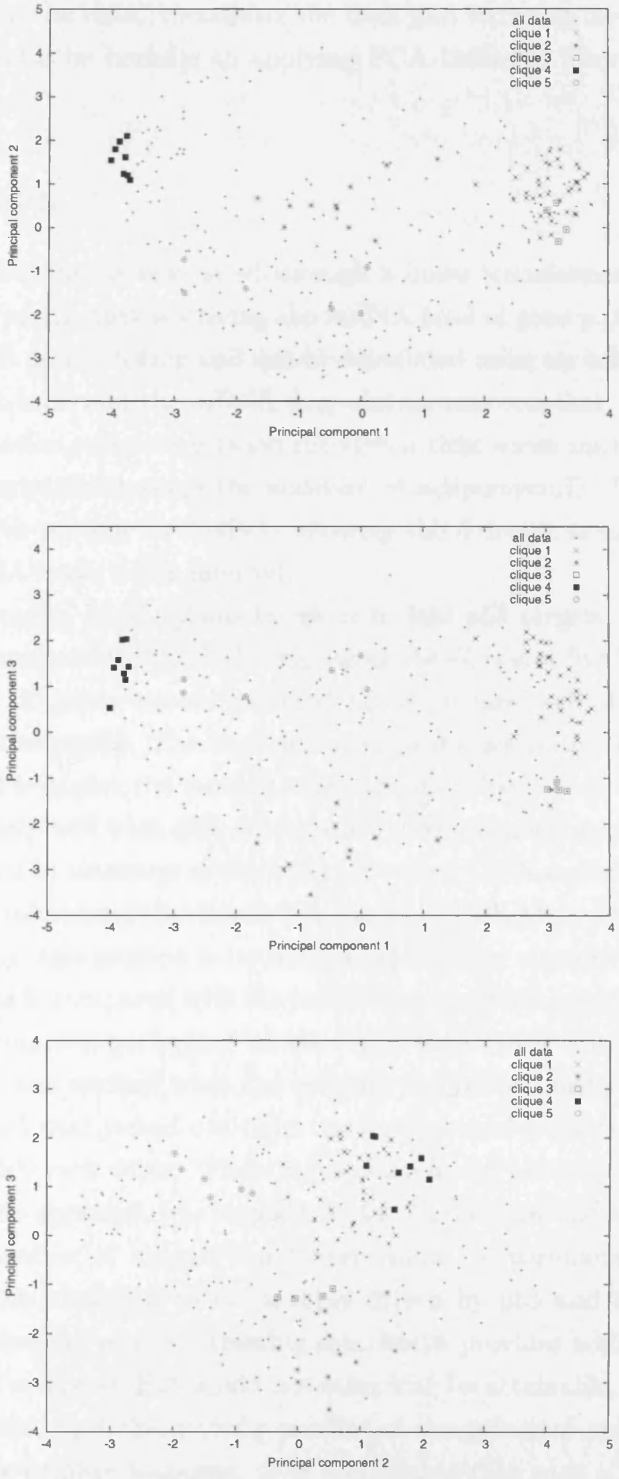


Figure 8.30: Plots of how the cliques found in section 8.6.3 distribute in the principal component space for the first three principal components.

the principal components do, even though the principal components can provide useful insights into the biological problem. PCA is good at determining the most significant factors that separate the data, visualising the data and reducing the dimensions of the space. There may also be benefits to applying PCA before or after the clique finding method.

8.8 Conclusion

$G_g(t)$ is a time profile that is associated through a linear transformation with the transcriptional activity profile that is driving the mRNA level of gene g. $G_g(t)$ is based on a simple model of gene transcription and can be calculated using an estimation of the rate of change of mRNA level and the mRNA degradation rate constant. A reasonable estimate of the degradation rates were found through a time series microarray experiment that stopped transcription through the addition of actinomycinD. The $G_g(t)$ connects the mRNA data with protein level effects allowing the dynamic transcription activities that drive the mRNA levels to be inferred.

An initial application of $G_g(t)$ was to use it to find p53 targets. This was done by producing a p53 representative profile by averaging the $G_g(t)$s of five known p53 targets. Then a ranked list of genes was constructed based on how well the genes correlated with the representative profile. This representative profile accurately followed the known active p53 profile. The higher the rank of a gene the more likely that it was a p53 target. This worked extremely well with 82% of the top 50 being verified as p53 targets, but the predictions decreased in accuracy as the rank increased. This method produced known as well as verified unknown p53 targets (for example PRKAB1, ASCC3L1 and P45). It was confirmed that this method is better than traditional clustering but only a small improvement was made compared with the correlation approach applied to mRNA levels.

K-means clustering was performed on the $G_g(t)$ data but it was found to be unsatisfactory because it was unclear what the optimal number of clusters were. Therefore, a method was devised that picked out tight clusters (merged cliques) of genes that were highly correlated with each other. These tight clusters can be used as training sets for other methods. This approach was applied to the $G_g(t)$ data and five merged cliques were found. The number of cliques found were robust to parameter change. Two of the training sets were identified to be possibly driven by p53 and the AP-1 complex. Not only does this method produce training sets, but it provides additional information about the biological response that would not otherwise be attainable. Most significantly it provides the number and the activity profiles of the principal mechanisms that are at work in the DNA damage response. It is remarkable that such a simple model that links the mRNA and protein levels could provide such a wealth of information. Principal component analysis was applied to the $G_g(t)$ data and there seems to be no relationship between the principal components and the principal activities. Despite this, principal

component analysis can be useful in examining the principal activities. In this case it provided evidence that one clique was co-regulated by p53 and NFκB.

In the future further research and experiments would need to be performed to identify more accurately what is behind the activities that are predicted to drive the DNA damage response. It would also be interesting to apply this method to DNA damage response data from cells that have received a smaller dose of radiation to see what activity profiles remain and which disappear, it is predicted that the AP-1 activity profile should not be observed. The use of $G_g(t)$ should be applicable to any biological system that produces a strong response in the cell and it would be interesting to use this technique on another system. A potential way to get further information from the $G_g(t)$ data would be decompose each gene's profile using the found training sets as composite components; this should allow the detection of co-regulated genes and allow the assignment of a probability on the membership of a gene to a particular transcription activity profile. It would also be interesting to see whether a model of negative regulation could be applied to detect negative regulation transcription profile activities. A simple model of negative regulation for gene g would be the following:

$$\frac{dx_g(t)}{dt} = \frac{B_g\theta_g}{\theta_g + f(t)} - D_g x_g(t), \tag{8.6}$$

where $f(t)$ is the regulator's concentration, B_g is the basal transcription rate, D_g is the degradation rate constant of gene g and θ_g controls how rapidly the transcription rate of production is suppressed. In this context, the calculated G profile will equal,

$$G_g(t) = \frac{B_g\theta_g}{\theta_g + f(t)},$$

therefore the inverse of the G time profile is,

$$G_g^{-1}(t) = \frac{1}{B_g} + \frac{1}{B_g\theta_g} f(t).$$

$G_g^{-1}(t)$ is therefore a linear transformation of the negative regulator's activity profile. It would be interesting to see whether the same tools could be used. There would be a problem about how to remove positive regulated genes, but it would be interesting to see if detectable activities emerged above the noise.

Chapter 9

Conclusion

p53 is the central protein in the DNA damage response and is part of a complex and extensive gene regulatory network. The p53 system has typically been studied qualitatively as a linear pathway, however this approach is insufficient to gain a full functional understanding of the dynamic nature of the network. The p53 network can be divided into two parts: the system that regulates the level of active p53 and the effect that p53 has on the rest of the cell. Previous to this work there had only been one mathematical model of the regulatory control of p53 and this was concerned with replicating the damped oscillations between p53 and MDM2 observed at the population level after DNA damage (Bar-Or *et al.*, 2000). Since then another model based on delay differential equations has successfully managed to produce damped oscillations (Monk, 2003a). The main focus though has been on the pulse dynamics observed at the single cell level (Lahav *et al.*, 2004), with two different models published very recently successfully replicated pulses in the p53 network (Ciliberto *et al.*, 2005; Ma *et al.*, 2005). There is still some doubt whether pulse dynamics occur as there is contrary experimental evidence (Jõers *et al.*, 2004). Once activated, p53 affects numerous subsystems and so global measurements are required. Microarray experiments are ideally suited to this. There has been no known use of mathematical models to extract p53 targets from microarray data apart from a recent paper produced by this group (Barenco *et al.*, 2005).

In this thesis various mathematical techniques have been applied and developed to examine the p53 gene regulatory network. The experimental system used examines the DNA damage response in the p53 wild type human lymphoid cell line MOLT 4. Cells were exposed to ionising radiation and time series measurements were gathered for mRNA and various proteins.

Ordinary differential equation models of p53 regulation at the protein level were proposed based on previous biological knowledge. These included "toy" models which suggested that the core of the network could produce oscillations at intermediary levels of DNA damage and that MOLT4 cells are more sensitive to DNA damage than MCF-7 cells. Parameter estimation is an essential technique in mathematical modelling. It quickly became apparent that established parameter estimation methods such as simulated annealing were not reliably converging to reasonable solutions with these models. This motivated the introduction of a new approach to parameter estimation based on linear algebra, collocation and B-splines. The spline acts as an intermediary between the model solution and the data, and is constructed so that it satisfies both to the greatest extent possible. The main constraint on this method is that it can only be applied to models that are linear in their parameters. This method always converged to a reasonable solution and produced very accurate results if there was a reasonable amount of data and the error in the data was small. Refinements were introduced to the method that force the spline to more closely represent the model, this radically improved the results when the error was higher. This parameter estimation method is fast, simple, and does not require the estimation of the initial conditions; it works well with the models proposed

in this thesis. If the data set is small it is very difficult to get reasonable parameters from any parameter estimation method. The same is true of the method proposed here, but various adaptations were made that improved the method's performance and made it usable. This method was applied to protein data gathered for this project. The results suggested that the ubiquitination of active p53 is an important mechanism in the system whereas MDM2 self-ubiquitination and p53 inactivation are not. Unfortunately it is impossible to give strength to these assertions due to the limited data set available.

Ordinary differential equation models were proposed that implemented mechanisms that controlled the localisation of p53; the activation of p53 which prevents export from the nucleus and the ubiquitination of p53 which allows the export of p53 from the nucleus. When a cell is damaged the ubiquitination is inhibited and the rate of activation is increased, thus confining p53 to the nucleus where it can be functionally active. Simulation of the models successfully reproduced the localisation of p53 with and without DNA damage observed in experiments. It was also shown that a faster and stronger response was achieved with no loss in recovery time by having the ubiquitination export mechanism. This is a possible reason why the two state p53 ubiquitination mechanism has developed. Finally, it was found that all the mechanisms that had their rate changed by the DNA damage signal (activation of p53, ubiquitination of p53 and inhibition of MDM2) contributed significantly to the performance of the response.

With a view to gain information about the protein level response to DNA damage using microarray mRNA data, especially downstream of p53, a quantity, $G_g(t)$, was proposed. This time profile correlates with the activity that drives the mRNA level of gene g. Using a relevant time profile of mRNA levels, and an estimate for both the mRNA degradation rate and the rate of change of mRNA level, $G_g(t)$ can be calculated using a basic model of gene transcription. As $G_g(t)$ is related to the transcription activities it has many possible applications. An initial application was to use it to find p53 targets using mRNA time profiles collected after DNA damage. By averaging the $G_g(t)$s of five known p53 targets a representative profile of p53 transcriptional activity was produced. This corresponded well with protein data. Using this representative profile a ranked list of potential p53 targets was produced by correlating $G_g(t)$ with the representative profile. Out of the top 50, a respectable 82% were verified as p53 targets. This method produces well known targets such as GADD45α, TP53 target gene 1, TRAF and cyclin G1, as well as verified unknown targets, for example PRKAB1, ASCC3L1 and P45. This method produces different targets to both clustering and correlation on mRNA time profiles.

Perhaps more interestingly, a method was devised that used the $G_g(t)$ profiles to pick out the principal transcription activities that drove the DNA damage response. This was based on picking out tight clusters of genes whose $G_g(t)$s were highly correlated. At the simplest level this method can be used to find groups of targets that share the same transcription factor and can be used in other methods as training sets, but these tight clusters also provides information about the biological response that would not otherwise

be attainable, such as the number and the activity profiles of the principal mechanisms that are at work in the DNA damage response. This method was applied to the DNA damage response time series and five training sets were found with associated principal activities. Two of these principal activities were associated with the p53 and possibly AP-1 transcription factors. The other three activities were less easily discerned but one might be co-regulation by p53 and NFκB, another seems to be closely related to cell cycle activity and the third associated with cell signalling.

9.1 The future

To gain a full functional understanding of the dynamic nature of the p53 network comprehensive and accurate data is required for both protein and mRNA. For protein in particular, the data is currently not sufficient for use with mathematical techniques, especially parameter estimation. In the next few years, with the development of new technologies such as protein arrays and computational advances, it will be possible to quantify the amount of protein accurately, quickly and reliably. Another important advancement will be the development of reliable and accurate single cell measurements. They provide much more information than population level measurements and are especially important when a cell population take two distinct directions, for example cell death or survival. Specifically, more single cell measurements are required on the p53 system to confirm whether there is a pulse-like response to DNA damage. Without data models are meaningless and so to advance the understanding of the DNA damage response a dual approach needs to be applied: improving the data and improving the models.

The entire p53 gene regulatory network is extremely complex with the possibility that many different mechanisms might significantly affect the response to DNA damage. Therefore the scope for modelling the system is large and here only basic models have been proposed. There are a number of areas that would be particularly interesting and important to model. Both p53 and MDM2 have homologs that appear to replicate only some of their functionality (Michael and Oren, 2002). These may provide a backup if the main proteins are faulty, but there is also speculation that there is a competitive effect which dampens down the effect the main protein has. For example, MDMX binds with p53 preventing MDM2 ubiquitinating p53, thus allowing it to survive longer (Stad *et al.*, 2000).

The levels of total p53 are suppressed through the AKT survival pathway (Testa and Bellacosa, 2001) and this pathway is triggered by intercellular survival signals (Lawlor and Alessi, 2001). It would be interesting to examine through modelling and experimentation whether inter-cell signalling affects the performance of the p53 network. It is proposed that if all the cells surrounding a single cell are damaged that the single cell is much more likely to commit to cell death. Another interesting question is whether p53 transcribes

genes that suppress survival signals.

p53 has a large number of binding sites for cofactors. These cofactors can have a dramatic effect on the activity of p53 and on which genes are targeted by p53 (Bode and Dong, 2004). It would be useful to examine how these cofactors are organised and how they achieve their effect. In particular what happens to the response of the network when there is competitive or co-operative binding effects between different groups of cofactors. Could the effects make the response more robust or heighten the p53 response?

Another interesting area to model is the effect that DNA damage has on active ATM. It has been assumed throughout this thesis that the amount of DNA damage is proportional to the amount of active ATM, but this might not be the case. It is possible that the levels of ATM activity are regulated in such a way that it is sustained even after the DNA damage has been repaired.

Improvements could be made to the models already proposed. In particular the localisation model could be made more realistic by, among other things, including MDM2's localisation properties, including more components, and a more complex description of both p53 activation and ubiquitination. It would be useful to perform experiments that tested some of the predictions made by the localisation model.

The parameter estimation technique introduced in this thesis could be further developed and analysed. In particular, it would be interesting to examine the effects of re-weighing the individual equations used in the algorithm. For example, re-weighing each data equation so that the relative error for the data points is consistent. Another possibility is that more weight needs to be placed on some parts of the spline, so that this part accurately describes the model to a greater degree. More work needs to be applied to the key question of how much data is required to produce reasonable estimates, and this is bound to depend to a certain extent on the amount of error in the data. It is clear from this thesis that much more protein data is required with a much better accuracy to produce reasonable parameter estimates.

Even though using $G_g(t)$ to discover the activity profiles has worked well, further research could be done on what the principal activity profiles actually are. This would require experiments, including possible knock-out experiments, to isolate the activities. Another interesting area that could be researched is what the principal activity profiles would be at different doses of DNA damage or different kinds of cell stress. In particular, at low doses it would be expected that the AP-1 profile would not be present. It would also be useful to apply this technique to a different system in the cell and test the generality of this approach, in particular the question of how strong a response is required before the activities are detectable needs to be tested. One possible system would be a T-cell's response after being activated. In this thesis only a couple of applications of $G_g(t)$ have been explored. It is likely that there are others, and of particular interest would be the decomposition of $G_g(t)$ in terms of the principal activity profiles. This would potentially lead to the discovery of co-regulated genes and make it easier to assign probabilities that

particular genes are targets of certain transcription factors. Even though only a simple model of gene transcription was used in this thesis it is possible that this approach could be applied to more complex models such as negative regulation.

References

R. T. Abraham and R. S. Tibbetts. Cell biology. Guiding ATM to broken DNA. *Science*, 308 (5721):510–511, Apr 2005.

F. S. Acton. *Numerical methods that work.* Mathematical Association of America, Washington, D.C., 1990.

Affymetrix. Statistical algorithms description document, 2002a.

Affymetrix. GeneChip expression analysis (data analysis fundamentals), 2002b.

B. Aguda. A quantitative analysis of the kinetics of the G(2) DNA damage checkpoint system. *Proc Natl Acad Sci U S A*, 96(20):11352–7, 1999.

B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular biology of the cell.* Garland Science, New York, 4th ed edition, 2002.

O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A*, 97(18):10101–10106, Aug 2000.

S. A. Amundson, A. Patterson, K. T. Do, and J. Fornace, A. J. A nucleotide excision repair master-switch: p53 regulated coordinate induction of global genomic repair genes. *Cancer Biol Ther*, 1(2):145–9, Mar-Apr 2002.

E. Appella and C. W. Anderson. Post-translational modifications and activation of p53 by genotoxic stresses. *Eur J Biochem*, 268(10):2764–72, May 2001.

C. J. Bakkenist and M. B. Kastan. DNA damage activates ATM through intermolecular autophosphorylation and dimer dissociation. *Nature*, 421(6922):499–506, Jan 2003.

P. Baldi and A. D. Long. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics*, 17(6):509–19, 2001.

E. Balint and K. Vousden. Activation and activities of the p53 tumour suppressor protein. *Br J Cancer*, 85(12):1813–23, 2001.

S. Banin, L. Moyal, S. Shieh, Y. Taya, C. W. Anderson, L. Chessa, N. I. Smorodinsky, C. Prives, Y. Reiss, Y. Shiloh, and Y. Ziv. Enhanced phosphorylation of p53 by ATM in response to DNA damage. *Science*, 281(5383):1674–1677, Sep 1998.

R. L. Bar-Or, R. Maya, L. A. Segel, U. Alon, A. J. Levine, and M. Oren. Generation of oscillations by the p53-Mdm2 feedback loop: a theoretical and experimental study. *Proc Natl Acad Sci U S A*, 97(21):11250–11255, Oct 2000.

Y. Barak, T. Juven, R. Haffner, and M. Oren. mdm2 expression is induced by wild type p53 activity. *EMBO J*, 12(2):461–8, 1993.

M. Barenco, D. Tomescu, D. Brewer, R. Callard, J. Stark, and M. Hubank. Prediction of transcription factor targets using hidden variable dynamic modelling. *Genome Biology (In Press)*, 2005.

S. Bates, S. Rowan, and K. H. Vousden. Characterisation of human cyclin G1 and G2: DNA damage inducible genes. *Oncogene*, 13(5):1103–9, Sep 5 1996.

D. Bell, J. Varley, T. Szydlo, D. Kang, D. Wahrer, K. Shannon, M. Lubratovich, S. Verselis, K. Isselbacher, J. Fraumeni, J. Birch, F. Li, J. Garber, and D. Haber. Heterozygous germ line

hCHK2 mutations in Li-Fraumeni syndrome. *Science*, 286(5449):2528–31, 1999.

D. Bergamaschi, Y. Samuels, N. J. O'Neil, G. Trigiante, T. Crook, J.-K. Hsieh, D. J. O'Connor, S. Zhong, I. Campargue, M. L. Tomlinson, P. E. Kuwabara, and X. Lu. iASPP oncoprotein is a key inhibitor of p53 conserved from worm to human. *Nat Genet*, 33(2):162–167, Feb 2003.

Bio-Rad. Quantity One 1-D analysis software (version 4.2.1) electronic help, 2005.

A. M. Bode and Z. Dong. Post-translational modification of p53 in tumorigenesis. *Nat Rev Cancer*, 4(10):793–805, Oct 2004.

V. Bours, V. Azarenko, E. Dejardin, and U. Siebenlist. Human RelB (I-Rel) functions as a kappa B site-dependent transactivating member of the family of Rel-related proteins. *Oncogene*, 9 (6):1699–1702, Jun 1994.

V. Bouvard, T. Zaitchouk, M. Vacher, A. Duthu, M. Canivet, C. Choisy-Rossi, M. Nieruchalski, and E. May. Tissue and cell-specific expression of the p53-target genes: bax, fas, mdm2 and waf1/p21, before and following ionising irradiation in mice. *Oncogene*, 19(5):649–660, Feb 2000.

M. Bouvet, L. Ellis, M. Nishizaki, T. Fujiwara, W. Liu, C. Bucana, B. Fang, J. Lee, and J. Roth. Adenovirus-mediated wild-type p53 gene transfer down-regulates vascular endothelial growth factor expression and inhibits angiogenesis in human colon cancer. *Cancer Res*, 58(11):2288–92, 1998.

S. Boyd, K. Tsai, and T. Jacks. An intact HDM2 RING-finger domain is required for nuclear exclusion of p53. *Nat Cell Biol*, 2(9):563–8, 2000.

R. F. Boyer. *Modern experimental biochemistry*. Benjamin Cummings, San Francisco; London, 3rd edition, 2000. ISBN 0805331115.

A. W. Braithwaite, J. A. Royds, and P. Jackson. The p53 story: layers of complexity. *Carcinogenesis*, 26(7):1161–1169, Jul 2005.

D. Brewer. Investigations of the p53 protein dna damage network using mathematical models. Master's thesis, Univeristy College London, 2002.

C. Bron and J. Kerbosch. Finding all cliques of an undirected graph [H]. *Communications Of The Acm*, 16(9):575–577, 1973.

F. M. Buffa, S. M. Bentzen, F. M. Daley, S. Dische, M. I. Saunders, P. I. Richman, and G. D. Wilson. Molecular marker profiles predict locoregional control of head and neck squamous cell carcinoma in a randomized trial of continuous hyperfractionated accelerated radiotherapy. *Clin Cancer Res*, 10(11):3745–54, Jun 1 2004.

F. Bunz, A. Dutriaux, C. Lengauer, T. Waldman, S. Zhou, J. P. Brown, J. M. Sedivy, K. W. Kinzler, and B. Vogelstein. Requirement for p53 and p21 to sustain G2 arrest after DNA damage. *Science*, 282(5393):1497–501, Nov 20 1998.

C. E. Canman, D. S. Lim, K. A. Cimprich, Y. Taya, K. Tamai, K. Sakaguchi, E. Appella, M. B. Kastan, and J. D. Siliciano. Activation of the ATM kinase by ionizing radiation and phosphorylation of p53. *Science*, 281(5383):1677–1679, Sep 1998.

M. F. Cardoso, R. L. Salcedo, and S. F. DeAzevedo. The simplex-simulated annealing approach to continuous non- linear optimization. *Computers & Chemical Engineering*, 20(9):1065–1080, 1996.

J. R. Cash and A. H. Karp. A variable order Runge-Kutta method for initial-value problems with rapidly varying right-hand sides. *ACM Transactions on Mathematical Software*, 16(3): 201–222, 1990.

C. Chao, S. Saito, J. Kang, C. Anderson, E. Appella, and Y. Xu. p53 transcriptional activity is essential for p53-dependent apoptosis following DNA damage. *EMBO J*, 19(18):4967–75, 2000.

N. Chehab, A. Malikzay, E. Stavridi, and T. Halazonetis. Phosphorylation of Ser-20 mediates stabilization of human p53 in response to DNA damage. *Proc Natl Acad Sci U S A*, 96(24): 13777–82, 1999.

N. H. Chehab, A. Malikzay, M. Appel, and T. D. Halazonetis. Chk2/hCds1 functions as a DNA damage checkpoint in G(1) by stabilizing p53. *Genes Dev*, 14(3):278–288, Feb 2000.

J. Chen, X. Wu, J. Lin, and A. J. Levine. mdm-2 inhibits the G1 arrest and apoptosis functions of the p53 tumor suppressor protein. *Mol Cell Biol*, 16(5):2445–2452, May 1996.

J. E. Chipuk, T. Kuwana, L. Bouchier-Hayes, N. M. Droin, D. D. Newmeyer, M. Schuler, and D. R. Green. Direct activation of Bax by p53 mediates mitochondrial membrane permeabilization and apoptosis. *Science*, 303(5660):1010–1014, Feb 2004.

J. E. Chipuk, L. Bouchier-Hayes, T. Kuwana, D. D. Newmeyer, and D. R. Green. PUMA couples the nuclear and cytoplasmic proapoptotic function of p53. *Science*, 309(5741):1732–1735, Sep 2005.

T. Chiu, D. Fang, J. Chen, Y. Wang, and C. Jeris. A robust and scalable clustering algorithm for mixed type attributes in large database environment. In *7th ACM SIGKDD Conference Proceedings*, page 263, 2001.

A. Ciliberto, B. Novak, and J. J. Tyson. Steady states and oscillations in the p53/Mdm2 network. *Cell Cycle*, 4(3):488–493, Mar 2005.

O. Cinquin. Fast-tracking morphogen diffusion. *J Theor Biol*, 238(3):532–40, Feb 2006.

A. R. Clarke, S. Gledhill, M. L. Hooper, C. C. Bird, and A. H. Wyllie. p53 dependence of early apoptotic and proliferative responses within the mouse intestinal epithelium following gamma-irradiation. *Oncogene*, 9(6):1767–1773, Jun 1994.

S. Dalziel. Numerical methods lecture notes. Web page, 1998. http://www.damtp.cam.ac.uk/user/fdl/people/sd/lectures/nummeth98/linear.htm.

K. Dameron, O. Volpert, M. Tainsky, and N. Bouck. Control of angiogenesis in fibroblasts by p53 regulation of thrombospondin-1. *Science*, 265(5178):1582–4, 1994.

D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis*, 1(2):224–227, 1979.

C. De Boor. *A practical guide to splines.* Applied mathematical sciences; 27. Springer, New York, 1978.

H. de Jong. Modeling and simulation of genetic regulatory systems: a literature review. *J Comput Biol*, 9(1):67–103, 2002.

G. Dennis, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol*, 4 (5):P3, 2003.

L. A. Donehower, M. Harvey, B. L. Slagle, M. J. McArthur, C. A. Montgomery, J. S. Butel, and A. Bradley. Mice deficient for p53 are developmentally normal but susceptible to spontaneous tumours. *Nature*, 356(6366):215–221, Mar 1992.

D. Dornan, I. Wertz, H. Shimizu, D. Arnott, G. D. Frantz, P. Dowd, K. O'Rourke, H. Koeppen, and V. M. Dixit. The ubiquitin ligase COP1 is a critical negative regulator of p53. *Nature*, 429(6987):86–92, May 2004.

J. C. Dunn. Well separated clusters and optimal fuzzy partitions. *J. Cybern.*, 4:95–104, 1974.

B. Efron, J. Storey, and R. Tibshirani. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96:1151–1160, 2001.

M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25):14863–8, 1998.

W. el Deiry, T. Tokino, V. Velculescu, D. Levy, R. Parsons, J. Trent, D. Lin, W. Mercer, K. Kinzler, and B. Vogelstein. WAF1, a potential mediator of p53 tumor suppression. *Cell*, 75(4): 817–25, 1993.

S. Erster and U. M. Moll. Stress-induced p53 runs a transcription-independent death program. *Biochem Biophys Res Commun*, 331(3):843–850, Jun 2005.

W. R. Esposito and C. A. Floudas. Global optimization for the parameter estimation of differential-algebraic systems. *Industrial & Engineering Chemistry Research*, 39(5):1291–1310, May 2000.

B. Everitt and G. Dunn. *Applied multivariate data analysis.* Hodder Arnold, 2001.

J. Falck, N. Mailand, R. G. S. sen, J. Bartek, and J. Lukas. The ATM-Chk2-Cdc25A checkpoint pathway guards against radioresistant DNA synthesis. *Nature*, 410(6830):842–847, Apr 2001.

J. Falck, J. Coates, and S. P. Jackson. Conserved modes of recruitment of ATM, ATR and DNA-PKcs to sites of DNA damage. *Nature*, 434(7033):605–611, Mar 2005.

C. P. Fall, E. S. Marland, J. M. Wagner, and J. J. Tyson, editors. *Computational cell biology.* Interdisciplinary applied mathematics; v. 20. Springer, New York, 2002. ISBN 0387953698 (alk. paper).

S. Fang, J. P. Jensen, R. L. Ludwig, K. H. Vousden, and A. M. Weissman. Mdm2 is a RING finger-dependent ubiquitin protein ligase for itself and p53. *J Biol Chem*, 275(12):8945–8951, Mar 2000.

E. Fehlberg. Classical fifth, sixth, seventh, and eighth order Runge-Kutta formulas with stepsize control. Technical Report TR R-287, NASA, 1968.

G. Filippini, S. Griffin, M. Uhr, H. Eppenberger, J. Bonilla, F. Cavalli, and G. Soldati. A novel flow cytometric method for the quantification of p53 gene expression. *Cytometry*, 31(3):180–6, Mar 1 1998.

M. Fiscella, H. Zhang, S. Fan, K. Sakaguchi, S. Shen, W. E. Mercer, G. F. Vande-Woude, P. M. O'Connor, and E. Appella. Wip1, a novel human protein phosphatase that is induced in response to ionizing radiation in a p53-dependent manner. *Proc Natl Acad Sci U S A*, 94(12): 6048–53, Jun 10 1997.

R. Fletcher and M. J. D. Powell. A rapidly convergent descent method for minimization. *The Computer Journal*, 6:163, 1963.

J. S. Fridman and S. W. Lowe. Control of apoptosis by p53. *Oncogene*, 22(56):9030–9040, Dec 2003.

A. Friedler, D. B. Veprintsev, L. O. Hansson, and A. R. Fersht. Kinetic instability of p53 core domain mutants: implications for rescue by small molecules. *J Biol Chem*, 278(26):24108–12, Jun 27 2003.

M. Fussenegger, J. Bailey, and J. Varner. A mathematical model of caspase function in apoptosis. *Nat Biotechnol*, 18(7):768–74, 2000.

R. R. Gabdoulline and R. C. Wade. Simulation of the diffusional association of barnase and barstar. *Biophys J*, 72(5):1917–29, May 1997.

N. A. Gershenfeld. *The nature of mathematical modeling.* Cambridge University Press, Cambridge, 1999.

R. Geyer, Z. Yu, and C. Maki. The MDM2 RING-finger domain is required to promote p53 nuclear export. *Nat Cell Biol*, 2(9):569–73, 2000.

W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov chain Monte Carlo in practice.* Chapman & Hall, London, 1996. ISBN 0412055511.

Z. Goldberg, R. V. Sionov, M. Berger, Y. Zwang, R. Perets, R. A. V. Etten, M. Oren, Y. Taya, and Y. Haupt. Tyrosine phosphorylation of Mdm2 by c-Abl: implications for p53 regulation. *EMBO J*, 21(14):3715–3727, Jul 2002.

G. H. Golub and J. M. Ortega. *Scientific computing and differential equations: an introduction to numerical methods.* Academic Press, Boston; London, 1992. ISBN 0122892550.

T. M. Gottlieb, J. F. M. Leal, R. Seger, Y. Taya, and M. Oren. Cross-talk between Akt, p53 and Mdm2: possible implications for the regulation of apoptosis. *Oncogene*, 21(8):1299–1303, Feb 2002.

J. Gu, L. Nie, D. Wiederschain, and Z. M. Yuan. Identification of p53 sequence elements that are required for MDM2-mediated nuclear export. *Mol Cell Biol*, 21(24):8533–8546, Dec 2001.

M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *Journal Of Intelligent Information Systems*, 17(2-3):107–145, 2001.

K. Harms, S. Nozell, and X. Chen. The common and distinct target genes of the p53 family transcription factors. *Cell Mol Life Sci*, 61(7-8):822–842, Apr 2004.

C. A. Harrington, C. Rosenow, and J. Retief. Monitoring gene expression using DNA microarrays. *Curr Opin Microbiol*, 3(3):285–91, 2000.

S. L. Harris and A. J. Levine. The p53 pathway: positive and negative feedback loops. *Oncogene*, 24(17):2899–2908, Apr 2005.

N. Hartsfield and G. Ringel. *Pearls in graph theory: a comprehensive introduction.* Academic Press, Boston; London, 1990. ISBN 0123285526.

S. Haupt, I. Louria-Hayon, and Y. Haupt. P53 licensed to kill? Operating the assassin. *J Cell Biochem*, 88(1):76–82, Jan 2003.

J. Hayakawa, S. Mittal, Y. Wang, K. S. Korkmaz, E. Adamson, C. English, M. Ohmichi, M. Omichi, M. McClelland, and D. Mercola. Identification of promoters bound by c-Jun/ATF2 during rapid large-scale gene activation following genotoxic stress. *Mol Cell*, 16(4):521–535, Nov 2004.

I. L. Hayon and Y. Haupt. p53: an internal investigation. *Cell Cycle*, 1(2):111–6, Mar-Apr 2002.

M. T. Heath. *Scientific computing: an introductory survey.* McGraw-Hill, New York; London, 1997. ISBN 0070276846 0071153365.

E. Hickman, M. Moroni, and K. Helin. The role of p53 and pRB in apoptosis and cancer. *Curr Opin Genet Dev*, 12(1):60–6, 2002.

N. S. Holter, M. Mitra, A. Maritan, M. Cieplak, J. R. Banavar, and N. V. Fedoroff. Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc Natl Acad Sci U S A*, 97(15):8409–8414, Jul 2000.

Y.-H. Huang, J. Y. Wu, Y. Zhang, and M. X. Wu. Synergistic and opposing regulation of the stress-responsive gene IEX-1 by p53, c-Myc, and multiple NF-kappaB/rel complexes. *Oncogene*, 21(44):6819–6828, Oct 2002.

T. R. Hupp and D. P. Lane. Allosteric activation of latent p53 tetramers. *Curr Biol*, 4(10): 865–875, Oct 1994.

B. J. Hwang, J. M. Ford, P. C. Hanawalt, and G. Chu. Expression of the p48 xeroderma pigmentosum gene is p53-dependent and is involved in global genomic repair. *Proc Natl Acad Sci U S A*, 96(2):424–8, Jan 19 1999.

H.-J. Im, M. R. Pittelkow, and R. Kumar. Divergent regulation of the growth-promoting gene IEX-1 by the p53 tumor suppressor and Sp1. *J Biol Chem*, 277(17):14612–14621, Apr 2002.

A. Ito, C. H. Lai, X. Zhao, S. Saito, M. H. Hamilton, E. Appella, and T. P. Yao. p300/CBP-mediated p53 acetylation is commonly induced by p53-activating agents and inhibited by MDM2. *EMBO J*, 20(6):1331–1340, Mar 2001.

G. Jimenez, M. Nister, J. Stommel, M. Beeche, E. Barcarse, X. Zhang, S. O'Gorman, and G. Wahl. A transactivation-deficient mouse model provides insights into Trp53 regulation and function. *Nat Genet*, 26(1):37–43, 2000.

A. Jõers, V. Jaks, J. Kase, and T. Maimets. p53-dependent transcription can exhibit both on/off and graded response after genotoxic stress. *Oncogene*, 23(37):6175–6185, Aug 2004.

S. Jones, A. Roe, L. Donehower, and A. Bradley. Rescue of embryonic lethality in Mdm2-deficient mice by absence of p53. *Nature*, 378(6553):206–8, 1995.

T. Kamijo, F. Zindy, M. Roussel, D. Quelle, J. Downing, R. Ashmun, G. Grosveld, and C. Sherr. Tumor suppression at the mouse INK4a locus mediated by the alternative reading frame product p19ARF. *Cell*, 91(5):649–59, 1997.

H. Kawai, D. Wiederschain, and Z. M. Yuan. Critical contribution of the MDM2 acidic domain to p53 ubiquitination. *Mol Cell Biol*, 23(14):4939–47, Jul 2003.

C. J. Kemp. You don't need a backbone to carry a tumour suppressor gene. *Nat Genet*, 21(2): 147–8, Feb 1999.

S. Khan, J. Moritsugu, and G. Wahl. Differential requirement for p19ARF in the p53-dependent arrest induced by DNA damage, microtubule disruption, and ribonucleotide depletion. *Proc Natl Acad Sci U S A*, 97(7):3266–71, 2000.

I. G. Kim, D. Y. Jun, U. Sohn, and Y. H. Kim. Cloning and expression of human mitotic centromere-associated kinesin gene. *Biochim Biophys Acta*, 1359(3):181–6, Dec 12 1997.

J.-W. Kim, H.-Y. Park, M.-J. Lee, M.-J. Jang, S.-Y. Lee, Y.-M. Park, D.-H. Son, Y.-C. Chang, Y.-S. Bae, and J.-Y. Kwak. Phosphatidic acid and tumor necrosis factor-alpha induce the expression of CD83 through mitogen activated protein kinase pathway in a CD34+ hematopoietic progenitor cell line, KG1. *Int Immunopharmacol*, 4(13):1603–1613, Dec 2004.

S. Kirkpatrick. Optimization by simulated annealing - quantitative studies. *Journal of Statistical Physics*, 34(5-6):975–986, 1984.

S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

K. W. Kohn and Y. Pommier. Molecular interaction map of the p53 and Mdm2 logic elements, which control the Off-On switch of p53 in response to DNA damage. *Biochem Biophys Res Commun*, 331(3):816–827, Jun 2005.

T. Kohonen. *Self-organizing maps*. Springer series in information sciences ; 30. Springer, Berlin ; London, 2nd edition, 1997.

E. Kreyszig. *Advanced engineering mathematics*. Wiley, New York; Chichester, 7th edition, 1993. ISBN 0471599891 (Wiley international edition) 0471553808 (cloth).

M. Kühne, E. Riballo, N. Rief, K. Rothkamm, P. A. Jeggo, and M. Löbrich. A double-strand break repair defect in ATM-deficient cells contributes to radiosensitivity. *Cancer Res*, 64(2): 500–508, Jan 2004.

V. Kvasnicka and J. Pospichal. A hybrid of simplex method and simulated annealing. *Chemometrics and Intelligent Laboratory Systems*, 39(2):161–173, 1997.

G. Lahav, N. Rosenfeld, A. Sigal, N. Geva-Zatorsky, A. J. Levine, M. B. Elowitz, and U. Alon. Dynamics of the p53-Mdm2 feedback loop in individual cells. *Nat Genet*, 36(2):147–150, Feb 2004.

Z. Lai, K. V. Ferry, M. A. Diamond, K. E. Wee, Y. B. Kim, J. Ma, T. Yang, P. A. Benfield, R. A. Copeland, and K. R. Auger. Human mdm2 mediates multiple mono-ubiquitination of p53 by a mechanism requiring enzyme isomerization. *J Biol Chem*, 276(33):31357–67, Aug 17 2001.

D. P. Lane. Cancer. p53, guardian of the genome. *Nature*, 358(6381):15–16, Jul 1992.

M. F. Lavin, G. Birrell, P. Chen, S. Kozlov, S. Scott, and N. Gueven. ATM signaling and genomic stability in response to DNA damage. *Mutat Res*, 569(1-2):123–132, Jan 2005.

M. A. Lawlor and D. R. Alessi. PKB/Akt: a key mediator of cell proliferation, survival and insulin responses? *J Cell Sci*, 114(Pt 16):2903–2910, Aug 2001.

J.-H. Lee and T. T. Paull. ATM activation by DNA double-strand breaks through the Mre11-Rad50-Nbs1 complex. *Science*, 308(5721):551–554, Apr 2005.

M. G. Lee, C. J. Norbury, N. K. Spurr, and P. Nurse. Regulated expression and phosphorylation of a possible mammalian cell-cycle control protein. *Nature*, 333(6174):676–9, Jun 16 1988.

R. P. Leng, Y. Lin, W. Ma, H. Wu, B. Lemmers, S. Chung, J. M. Parant, G. Lozano, R. Hakem, and S. Benchimol. Pirh2, a p53-induced ubiquitin-protein ligase, promotes p53 degradation. *Cell*, 112(6):779–791, Mar 2003.

J. I.-J. Leu, P. Dumont, M. Hafey, M. E. Murphy, and D. L. George. Mitochondrial p53 activates Bak and causes disruption of a Bak-Mcl1 complex. *Nat Cell Biol*, 6(5):443–450, May 2004.

C. Q. Li, A. I. Robles, C. L. Hanigan, L. J. Hofseth, L. J. Trudel, C. C. Harris, and G. N. Wogan. Apoptotic signaling pathways induced by nitric oxide in human lymphoblastoid cells expressing wild-type or mutant p53. *Cancer Res*, 64(9):3022–9, May 1 2004a.

M. Li, D. Chen, A. Shiloh, J. Luo, A. Y. Nikolaev, J. Qin, and W. Gu. Deubiquitination of p53 by HAUSP is an important pathway for p53 stabilization. *Nature*, 416(6881):648–653, Apr

2002.

M. Li, C. L. Brooks, F. Wu-Baer, D. Chen, R. Baer, and W. Gu. Mono- versus polyubiquitination: differential control of p53 fate by Mdm2. *Science*, 302(5652):1972–5, Dec 12 2003.

M. Li, C. L. Brooks, N. Kon, and W. Gu. A dynamic role of HAUSP in the p53-Mdm2 pathway. *Mol Cell*, 13(6):879–886, Mar 2004b.

S. H. Liang and M. F. Clarke. Regulation of p53 localization. *Eur J Biochem*, 268(10):2779–83, May 2001.

S.-K. Lim, J.-M. Shin, Y.-S. Kim, and K.-H. Baek. Identification and characterization of murine mHAUSP encoding a deubiquitinating enzyme that regulates the status of p53 ubiquitination. *Int J Oncol*, 24(2):357–364, Feb 2004.

W. Lin, F. Lin, and J. Nevins. Selective induction of E2F1 in response to DNA damage, mediated by ATM-dependent phosphorylation. *Genes Dev*, 15(14):1833–44, 2001.

G. Liu and X. Chen. The ferredoxin reductase gene is regulated by the p53 family and sensitizes cells to oxidative stress-induced apoptosis. *Oncogene*, 21(47):7195–204, Oct 17 2002.

S. W. Lowe and C. J. Sherr. Tumor suppression by Ink4a-Arf: progress and puzzles. *Curr Opin Genet Dev*, 13(1):77–83, Feb 2003.

J. Luo, F. Su, D. Chen, A. Shiloh, and W. Gu. Deacetylation of p53 modulates its effect on cell growth and apoptosis. *Nature*, 408(6810):377–381, Nov 2000.

L. Ma, J. Wagner, J. J. Rice, W. Hu, A. J. Levine, and G. A. Stolovitzky. A plausible model for the digital response of p53 to DNA damage. *Proc Natl Acad Sci U S A*, 102(40):14266–14271, Oct 2005.

P. F. Macgregor and J. A. Squire. Application of microarrays to the analysis of gene expression in cancer. *Clin Chem*, 48(8):1170–7, 2002.

J. Maddox. Is molecular biology yet a science? *Nature*, 355(6357):201, Jan 1992.

C. G. Maki and P. M. Howley. Ubiquitination of p53 and p21 is differentially affected by ionizing and UV radiation. *Mol Cell Biol*, 17(1):355–63, Jan 1997.

J. Mao, K. Lindsay, A. Balmain, and T. Wheldon. Stochastic modelling of tumorigenesis in p53 deficient mice. *Br J Cancer*, 77(2):243–52, 1998.

J. Mao, K. Lindsay, R. Mairs, and T. Wheldon. The effect of tissue-specific growth patterns of target stem cells on the spectrum of tumours resulting from multistage tumorigenesis. *J Theor Biol*, 210(1):93–100, 2001.

J. H. Mao, J. Perez-Losada, D. Wu, R. Delrosario, R. Tsunematsu, K. I. Nakayama, K. Brown, S. Bryson, and A. Balmain. Fbxw7/Cdc4 is a p53-dependent, haploinsufficient tumour suppressor gene. *Nature*, 432(7018):775–9, Dec 9 2004.

N. F. Marko, P. B. Dieffenbach, G. Yan, S. Ceryak, R. W. Howell, T. A. McCaffrey, and V. W. Hu. Does metabolic radiolabeling stimulate the stress response? Gene expression profiling reveals differential cellular responses to internal beta vs. external gamma radiation. *Faseb J*, 17(11):1470–86, Aug 2003.

D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *SIAM J. Appl. Math.*, 11:431–441, 1963.

S. Mason, O. Loughran, and N. La Thangue. p14(ARF) regulates E2F activity. *Oncogene*, 21 (27):4220–30, 2002.

C. K. Mathews, K. E. Van Holde, and K. G. Ahern. *Biochemistry*. Benjamin Cummings, San Francisco, Calif.; Harlow, 3rd edition, 2000. ISBN 0805330666: No price.

R. Maya, M. Balass, S. Kim, D. Shkedy, J. Leal, O. Shifman, M. Moas, T. Buschmann, Z. Ronai, Y. Shiloh, M. Kastan, E. Katzir, and M. Oren. ATM-dependent phosphorylation of Mdm2 on serine 395: role in p53 activation by DNA damage. *Genes Dev*, 15(9):1067–77, 2001.

L. D. Mayo and D. B. Donner. The PTEN, Mdm2, p53 tumor suppressor-oncoprotein network. *Trends Biochem Sci*, 27(9):462–467, Sep 2002.

A. J. Merritt, C. S. Potten, C. J. Kemp, J. A. Hickman, A. Balmain, D. P. Lane, and P. A. Hall. The role of p53 in spontaneous and radiation-induced apoptosis in the gastrointestinal tract of normal and p53-deficient mice. *Cancer Res*, 54(3):614–617, Feb 1994.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machine. *J. Chem. Phys.*, 21:1087–1091, 1953.

D. Michael and M. Oren. The p53 and Mdm2 families in cancer. *Curr Opin Genet Dev*, 12(1): 53–59, Feb 2002.

D. Michael and M. Oren. The p53-Mdm2 module and the ubiquitin system. *Semin Cancer Biol*, 13(1):49–58, Feb 2003.

M. Mihara, S. Erster, A. Zaika, O. Petrenko, T. Chittenden, P. Pancoska, and U. M. Moll. p53 has a direct apoptogenic role at the mitochondria. *Mol Cell*, 11(3):577–590, Mar 2003.

N. A. Monk. Oscillatory expression of Hes1, p53, and NF-kappaB driven by transcriptional time delays. *Curr Biol*, 13(16):1409–13, 2003a.

N. A. Monk. Oscillatory expression of Hes1, p53, and NF-kappaB driven by transcriptional time delays. *Curr Biol*, 13(16), 2003b. Supplementary Material.

J. A. Nelder and R. Mead. A simplex method for function minimization. *The computer journal*, 7:308–313, 1965.

C. J. Norbury and B. Zhivotovsky. DNA damage-induced apoptosis. *Oncogene*, 23(16):2797–2808, Apr 2004.

S. H. Northrup and H. P. Erickson. Kinetics of protein-protein association explained by Brownian dynamics computer simulation. *Proc Natl Acad Sci U S A*, 89(8):3338–42, Apr 15 1992.

B. Novak and J. Tyson. Numerical analysis of a comprehensive model of M-phase control in Xenopus oocyte extracts and intact embryos. *J Cell Sci*, 106 (Pt 4):1153–68, 1993.

B. Novak, A. Csikasz-Nagy, B. Gyorffy, K. Chen, and J. Tyson. Mathematical model of the fission yeast cell cycle with checkpoint controls at the G1/S, G2/M and metaphase/anaphase transitions. *Biophys Chem*, 72(1-2):185–200, 1998.

S. Obad, H. Brunnstrom, J. Vallon-Christersson, A. Borg, K. Drott, and U. Gullberg. Staf50 is a novel p53 target gene conferring reduced clonogenic growth of leukemic U-937 cells. *Oncogene*, 23(23):4050–9, May 20 2004.

A. O'Brate and P. Giannakakou. The importance of p53 location: nuclear or cytoplasmic zip code? *Drug Resist Updat*, 6(6):313–22, Dec 2003.

H. Offer, I. Zurer, G. Banfalvi, M. Reha'k, A. Falcovitz, M. Milyavsky, N. Goldfinger, and V. Rotter. p53 modulates base excision repair activity in a cell cycle-specific manner after genotoxic stress. *Cancer Res*, 61(1):88–96, 2001.

K. Okamoto, H. Li, M. R. Jensen, T. Zhang, Y. Taya, S. S. Thorgeirsson, and C. Prives. Cyclin G recruits PP2A to dephosphorylate Mdm2. *Mol Cell*, 9(4):761–771, Apr 2002.

J. Oliner, K. Kinzler, P. Meltzer, D. George, and B. Vogelstein. Amplification of a gene encoding a p53-associated protein in human sarcomas. *Nature*, 358(6381):80–3, 1992.

J. Parant, A. Chavez-Reyes, N. A. Little, W. Yan, V. Reinke, A. G. Jochemsen, and G. Lozano. Rescue of embryonic lethality in Mdm4-null mice by loss of Trp53 suggests a nonoverlapping pathway with MDM2 to regulate p53. *Nat Genet*, 29(1):92–95, Sep 2001.

W. R. Park and Y. Nakamura. p53CSV, a novel p53-inducible gene involved in the p53-dependent cell-survival pathway. *Cancer Res*, 65(4):1197–206, Feb 15 2005.

A. Pietzsch, C. Büchler, and G. Schmitz. Genomic organization, promoter cloning, and chromosomal localization of the Dif-2 gene. *Biochem Biophys Res Commun*, 245(3):651–657, Apr 1998.

M. J. D. Powell. An iterative method for finding stationary value of a function of several variables. *The Computer Journal*, 5:147, 1962.

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes in C++:*

the art of scientific computing. Cambridge University Press, Cambridge, 2nd edition, 2002.

J. Quackenbush. Computational analysis of microarray data. *Nat Rev Genet*, 2(6):418–27, 2001.

Y. F. Ramos, R. Stad, J. Attema, L. T. Peltenburg, A. J. van der Eb, and A. G. Jochemsen. Aberrant expression of HDMX proteins in tumor cells correlates with wild-type p53. *Cancer Res*, 61(5):1839–1842, Mar 2001.

N. C. Reich, M. Oren, and A. J. Levine. Two distinct mechanisms regulate the levels of a cellular tumor antigen, p53. *Mol Cell Biol*, 3(12):2143–50, Dec 1983.

J. A. Rice. *Mathematical statistics and data analysis.* Duxbury Press, Belmont, Calif, 2nd edition, 1995. ISBN 0534209343.

T. Rich, C. Watson, and A. Wyllie. Apoptosis: the germs of death. *Nat Cell Biol*, 1(3):E69–71, 1999.

T. Rich, R. Allen, and A. Wyllie. Defying death after DNA damage. *Nature*, 407(6805):777–83, 2000.

E. P. Rogakou, D. R. Pilch, A. H. Orr, V. S. Ivanova, and W. M. Bonner. DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139. *J Biol Chem*, 273(10):5858–5868, Mar 1998.

H. Rosenbrock. An automatic method for finding the greatest or least value of a function. *The Computer Journal*, 3:175, 1960.

K. Rothkamm and M. Löbrich. Evidence for a lack of DNA double-strand break repair in human cells exposed to very low x-ray doses. *Proc Natl Acad Sci U S A*, 100(9):5057–5062, Apr 2003.

J. P. Rouault, N. Falette, F. Guehenneux, C. Guillot, R. Rimokh, Q. Wang, C. Berthet, C. Moyret-Lalle, P. Savatier, B. Pain, P. Shaw, R. Berger, J. Samarut, J. P. Magaud, M. Ozturk, C. Samarut, and A. Puisieux. Identification of BTG2, an antiproliferative p53-dependent component of the DNA damage cellular response pathway. *Nat Genet*, 14(4):482–6, Dec 1996.

J. J. Ryan, E. Prochownik, C. A. Gottlieb, I. J. Apel, R. Merino, G. Nunez, and M. F. Clarke. c-myc and bcl-2 modulate p53 function by altering p53 subcellular trafficking during the cell cycle. *Proc Natl Acad Sci U S A*, 91(13):5878–82, Jun 21 1994.

R. P. Ryseck and R. Bravo. c-JUN, JUN B, and JUN D differ in their binding affinities to AP-1 and CRE consensus sequences: effect of FOS proteins. *Oncogene*, 6(4):533–42, Apr 1991.

Y. Samuels-Lev, D. J. O'Connor, D. Bergamaschi, G. Trigiante, J. K. Hsieh, S. Zhong, I. Campargue, L. Naumovski, T. Crook, and X. Lu. ASPP proteins specifically stimulate the apoptotic function of p53. *Mol Cell*, 8(4):781–794, Oct 2001.

K. Savitsky, A. Bar-Shira, S. Gilad, G. Rotman, Y. Ziv, L. Vanagaite, D. A. Tagle, S. Smith, T. Uziel, and S. Sfez. A single ataxia telangiectasia gene with a product similar to PI-3 kinase. *Science*, 268(5218):1749–1753, Jun 1995.

M. Schuler and D. R. Green. Transcription, apoptosis and p53: catch-22. *Trends Genet*, 21(3): 182–187, Mar 2005.

A. Schulze and J. Downward. Navigating gene expression using microarrays–a technology review. *Nat Cell Biol*, 3(8):E190–5, 2001.

P. Sebastiani and M. F. Ramoni. Statistical challenges in functional genomics. Technical report, Children's Hospital Informatics Program, Harvard Medical School and the Department of Mathematics and Statistics, University of Massachusetts at Amherst., Feb 2002.

O. A. Sedelnikova, E. P. Rogakou, I. G. Panyutin, and W. M. Bonner. Quantitative detection of (125)IdU-induced DNA double-strand breaks with gamma-H2AX antibody. *Radiat Res*, 158 (4):486–492, Oct 2002.

E. Shaulian and M. Karin. Ap-1 as a regulator of cell life and death. *Nat Cell Biol*, 4(5):E131–6, May 2002.

G. Shaulsky, N. Goldfinger, A. Ben-Ze'ev, and V. Rotter. Nuclear accumulation of p53 protein is mediated by several nuclear localization signals and plays a role in tumorigenesis. *Mol Cell Biol*, 10(12):6565–77, Dec 1990.

G. Sherlock. Analysis of large-scale gene expression data. *Curr Opin Immunol*, 12(2):201–5, 2000.

C. Sherr and J. Weber. The ARF/p53 pathway. *Curr Opin Genet Dev*, 10(1):94–9, 2000.

A. Shmueli and M. Oren. Regulation of p53 by Mdm2: fate is in the numbers. *Mol Cell*, 13(1): 4–5, Jan 16 2004.

J. L. D. Sigala, V. Bottero, D. B. Young, A. Shevchenko, F. Mercurio, and I. M. Verma. Activation of transcription factor NF-kappaB requires ELKS, an IkappaB kinase regulatory subunit. *Science*, 304(5679):1963–1967, Jun 2004.

E. A. Slee, D. J. O'Connor, and X. Lu. To die or not to die: how does p53 decide? *Oncogene*, 23 (16):2809–2818, Apr 2004.

G. Smith, R. Cary, N. Lakin, B. Hann, S. Teo, D. Chen, and S. Jackson. Purification and DNA binding properties of the ataxia-telangiectasia gene product ATM. *Proc Natl Acad Sci U S A*, 96(20):11134–9, 1999.

M. L. Smith, I. T. Chen, Q. Zhan, I. Bae, C. Y. Chen, T. M. Gilmer, M. B. Kastan, P. M. O'Connor, and J. Fornace, A. J. Interaction of the p53-regulated protein Gadd45 with proliferating cell nuclear antigen. *Science*, 266(5189):1376–80, Nov 25 1994.

T. Soussi, K. Dehouche, and C. Beroud. p53 website and analysis of p53 gene mutations in human cancer: forging a link between epidemiology and carcinogenesis. *Hum Mutat*, 15(1): 105–13, 2000.

R. Stad, Y. F. Ramos, N. Little, S. Grivell, J. Attema, A. J. van Der Eb, and A. G. Jochemsen. Hdmx stabilizes Mdm2 and p53. *J Biol Chem*, 275(36):28039–28044, Sep 2000.

J. Stark, D. Brewer, M. Barenco, D. Tomescu, R. Callard, and M. Hubank. Reconstructing gene networks: what are the limits? *Biochem Soc Trans*, 31(Pt 6):1519–25, 2003.

C. Stevens, L. Smith, and N. B. L. Thangue. Chk2 activates E2F-1 in response to DNA damage. *Nat Cell Biol*, 5(5):401–409, May 2003.

J. M. Stommel, N. D. Marchenko, G. S. Jimenez, U. M. Moll, T. J. Hope, and G. M. Wahl. A leucine-rich nuclear export signal in the p53 tetramerization domain: regulation of subcellular localization and p53 activity by NES masking. *EMBO J*, 18(6):1660–1672, Mar 1999.

W. J. H. Stortelder. Parameter estimation in dynamic systems. *Mathematics and Computers in Simulation*, 42(2-3):135–142, 1996.

S. H. Strogatz. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering.* Studies in nonlinearity. Perseus Books, Cambridge, Mass., 2000.

B. Talcott and M. S. Moore. Getting across the nuclear pore complex. *Trends Cell Biol*, 9(8): 312–8, Aug 1999.

H. Tanaka, H. Arakawa, T. Yamaguchi, K. Shiraishi, S. Fukuda, K. Matsui, Y. Takei, and Y. Nakamura. A ribonucleotide reductase gene involved in a p53-dependent cell-cycle checkpoint for DNA damage. *Nature*, 404(6773):42–9, 2000.

W. Tao and A. J. Levine. P19(ARF) stabilizes p53 by blocking nucleo-cytoplasmic shuttling of Mdm2. *Proc Natl Acad Sci U S A*, 96(12):6937–41, Jun 8 1999.

G. Teoh, M. Urashima, A. Ogata, D. Chauhan, J. A. DeCaprio, S. P. Treon, R. L. Schlossman, and K. C. Anderson. MDM2 protein overexpression promotes proliferation and survival of multiple myeloma cells. *Blood*, 90(5):1982–92, Sep 1 1997.

J. R. Testa and A. Bellacosa. AKT plays a central role in tumorigenesis. *Proc Natl Acad Sci U S A*, 98(20):10983–10985, Sep 2001.

R. Thomas and M. Kaufman. Multistationarity, the basis of cell differentiation and memory. I. Structural conditions of multistationarity and other nontrivial behavior. *Chaos*, 11(1):170–179, Mar 2001.

D. Tieu, W. R. Cluett, and A. Penlidis. A comparison of collocation methods for solving dynamic optimization problems. *Computers & Chemical Engineering*, 19(4):375–381, Apr 1995.

I. B. Tjoa and L. T. Biegler. Simultaneous solution and optimization strategies for parameter-

estimation of differential-algebraic equation systems. *Industrial & Engineering Chemistry Research*, 30(2):376–385, Feb 1991.

J. Tomfohr, J. Lu, and T. B. Kepler. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, 6:225, Sep 2005.

F. M. Torres, E. Agichtein, L. Grinberg, G. W. Yu, and R. Q. Topper. A note on the application of the "Boltzmann simplex"-simulated annealing algorithm to global optimizations of argon and water clusters. *Theochem-Journal of Molecular Structure*, 419:85–95, 1997.

V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, 98(9):5116–21, 2001.

J. Tyson. Models of cell cycle control in eukaryotes. *J Biotechnol*, 71(1-3):239–44, 1999.

J. Tyson and B. Novak. Regulation of the eukaryotic cell cycle: molecular antagonism, hysteresis, and irreversible transitions. *J Theor Biol*, 210(2):249–63, 2001.

B. Van den Bosch and L. Hellinck. New method for estimation of parameters in differential equations. *Aiche Journal*, 20(2):250–256, 1974.

N. L. van Hal, O. Vorst, A. M. van Houwelingen, E. J. Kok, A. Peijnenburg, A. Aharoni, A. J. van Tunen, and J. Keijer. The application of DNA microarrays in gene expression analysis. *J Biotechnol*, 78(3):271–80, 2000.

S. Varmeh-Ziaie, I. Okan, Y. Wang, K. P. Magnusson, P. Warthoe, M. Strauss, and K. G. Wiman. Wig-1, a new p53-induced gene encoding a zinc finger protein. *Oncogene*, 15(22):2699–704, Nov 27 1997.

S. Velasco-Miguel, L. Buckbinder, P. Jean, L. Gelbert, R. Talbott, J. Laidlaw, B. Seizinger, and N. Kley. PA26, a novel target of the p53 tumor suppressor and member of the GADD family of DNA damage and growth arrest inducible genes. *Oncogene*, 18(1):127–37, Jan 7 1999.

M. P. Vierboom, S. Zwaveling, G. M. J. Bos, M. Ooms, G. M. Krietemeijer, C. J. Melief, and R. Offringa. High steady-state levels of p53 are not a prerequisite for tumor eradication by wild-type p53-specific cytotoxic T lymphocytes. *Cancer Res*, 60(19):5508–13, Oct 1 2000.

B. Vogelstein, D. Lane, and A. Levine. Surfing the p53 network. *Nature*, 408(6810):307–10, 2000.

K. H. Vousden. p53: death star. *Cell*, 103(5):691–4, Nov 22 2000.

K. H. Vousden. Apoptosis. p53 and PUMA: a deadly duo. *Science*, 309(5741):1685–1686, Sep 2005.

K. H. Vousden and G. F. Woude. The ins and outs of p53. *Nat Cell Biol*, 2(10):E178–80, Oct 2000.

G. Wahl and A. Carr. The evolution of diverse biological responses to DNA damage: insights from yeast and p53. *Nat Cell Biol*, 3(12):E277–86, 2001.

G. Wahl, S. Linke, T. Paulson, and L. Huang. Maintaining genetic stability through TP53 mediated checkpoint control. *Cancer Surv*, 29:183–219, 1997.

M. E. Wall, P. A. Dyck, and T. S. Brettin. SVDMAN–singular value decomposition analysis of microarray data. *Bioinformatics*, 17(6):566–568, Jun 2001.

M. E. Wall, A. Rechtsteiner, and L. M. Rocha. Singular value decomposition and principal component analysis. In D. P. Berrar, W. Dubitzky, and M. Granzow, editors, *A practical approach to microarray data analysis*, chapter 5, pages 91–109. Kluwer Academic Publishers, 2002.

F. S. Wang. A modified collocation method for solving differential-algebraic equations. *Applied Mathematics And Computation*, 116(3):257–278, Dec 2000.

M. Wani, Q. Zhu, M. El-Mahdy, and A. Wani. Influence of p53 tumor suppressor protein on bias of DNA repair and apoptotic response in human cells. *Carcinogenesis*, 20(5):765–72, 1999.

J. D. Weber, L. J. Taylor, M. F. Roussel, C. J. Sherr, and D. Bar-Sagi. Nucleolar Arf sequesters Mdm2 and activates p53. *Nat Cell Biol*, 1(1):20–6, May 1999. 1465-7392 Journal Article.

G. S. Wu, T. F. Burns, r. McDonald, E. R., W. Jiang, R. Meng, I. D. Krantz, G. Kao, D. D.

Gan, J. Y. Zhou, R. Muschel, S. R. Hamilton, N. B. Spinner, S. Markowitz, G. Wu, and W. S. el Deiry. KILLER/DR5 is a DNA damage-inducible p53-regulated death receptor gene. *Nat Genet*, 17(2):141–3, Oct 1997.

A. Yang, R. Schweitzer, D. Sun, M. Kaghad, N. Walker, R. T. Bronson, C. Tabin, A. Sharpe, D. Caput, C. Crum, and F. McKeon. p63 is essential for regenerative proliferation in limb, craniofacial and epithelial development. *Nature*, 398(6729):714–718, Apr 1999.

A. Yang, N. Walker, R. Bronson, M. Kaghad, M. Oosterwegel, J. Bonnin, C. Vagner, H. Bonnet, P. Dikkes, A. Sharpe, F. McKeon, and D. Caput. p73-deficient mice have neurological, pheromonal and inflammatory defects but lack spontaneous tumours. *Nature*, 404(6773):99–103, Mar 2000.

O. Yazgan and C. M. Pfarr. Regulation of two JunD isoforms by Jun N-terminal kinases. *J Biol Chem*, 277(33):29710–29718, Aug 2002.

K. S. Yee and K. H. Vousden. Complicating the complexity of p53. *Carcinogenesis*, 26(8): 1317–1322, Aug 2005.

K. Y. Yeung and W. L. Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–774, Sep 2001.

M. K. S. Yeung, J. Tegnér, and J. J. Collins. Reverse engineering gene networks using singular value decomposition and robust regression. *Proc Natl Acad Sci U S A*, 99(9):6163–6168, Apr 2002.

E. Yonish-Rouach, D. Resnitzky, J. Lotem, L. Sachs, A. Kimchi, and M. Oren. Wild-type p53 induces apoptosis of myeloid leukaemic cells that is inhibited by interleukin-6. *Nature*, 352 (6333):345–347, Jul 1991.

D. Yu, T. Jing, B. Liu, J. Yao, M. Tan, T. J. McDonnell, and M. C. Hung. Overexpression of ErbB2 blocks Taxol-induced apoptosis by upregulation of p21Cip1, which inhibits p34Cdc2 kinase. *Mol Cell*, 2(5):581–91, Nov 1998.

Y. Zhang and Y. Xiong. A p53 amino-terminal nuclear export signal inhibited by DNA damage-induced phosphorylation. *Science*, 292(5523):1910–5, Jun 8 2001. 0036-8075 Journal Article.

J. Zhu, D. DeRyckere, F. Li, Y. Wan, and J. DeGregori. A role for E2F1 in the induction of ARF, p53, and apoptosis during thymic negative selection. *Cell Growth Differ*, 10(12):829–38, 1999.

W. Zou, X. Liu, P. Yue, Z. Zhou, M. B. Sporn, R. Lotan, F. R. Khuri, and S.-Y. Sun. c-Jun NH2-terminal kinase-mediated up-regulation of death receptor 5 contributes to induction of apoptosis by the novel synthetic triterpenoid methyl-2-cyano-3,12-dioxooleana-1, 9-dien-28-oate in human lung cancer cells. *Cancer Res*, 64(20):7570–7578, Oct 2004.

Appendix A

Experimental techniques

A.1 Western blots

Western blotting or immunoblotting is the standard technique for detecting specific proteins in a mixture (Mathews *et al.*, 2000). The first step is to extract the protein (Trizol, Invitrogen) and then apply SDS polyacrylamide-gel electrophoresis (SDS-PAGE). SDS-PAGE separates the proteins according to weight. The protein extract is placed in lanes at the top of a highly cross-linked gel of polyacrylamide and a current is applied across the gel, which attracts proteins at different speeds depending on their size. Once the proteins are separated the protein is transferred to a nitrocellulose membrane to allow the rest of the steps to be performed easily (Boyer, 2000). The membrane binds protein non-specifically so the membrane is blocked by placing the membrane in among other things, non-fat dry milk.

Next the membrane is probed for the protein of interest, which is a two step process. Firstly the primary antibody is incubated with the membrane for about half an hour, in a solution which contains water, salt, a small amount of buffer to keep the solution near neutral pH and some protein which helps non-specific binding. The primary antibody binds to the protein of interest. The affinity that the antibody has with the target protein seriously affects the quality of the results. After the incubation period the unbound antibody is washed away. In the second step the membrane is exposed to a secondary antibody that binds to the primary antibody. The secondary antibody is also linked to an enzyme tag that allows the visual identification of where the membrane is bound and hence where the protein is. There are many detection techniques but here enhanced chemiluminescence will be used. This relies on an enzyme that generates light when it interacts with the secondary antibody. The membrane is immersed in this enzyme and light is captured by exposing the membrane to film. This gives a permanent record of the presence or absence of the protein of interest.

A.2 QPCR

QPCR is a method to quantify the amount of mRNA present through the Polymerase Chain Reaction (PCR) technique. PCR amplifies the starting amount of DNA, doubles the amount of DNA present. A PCR cycle consists of three steps, in the first step the DNA is heated to separate the DNA strands. In the second step temperature is reduced so that primers (short, artificial DNA strands) can anneal to the complementary binding sites of the DNA being reproduced. This allows DNA-Polymerase, in the third step, to extend the primers and synthesise a new complementary DNA strand. In QPCR, there is a fluorescent marker in the primer that fluoresces once the polymerase has been displaced. The amount of fluorescence given off gives a measure of how much DNA is produced in that cycle. Initially, the amount of fluorescence is too low to be distinguished from noise. In each cycle the amount of DNA doubles so the number of cycles it takes to reach some fixed amount of fluorescence provides a measure of the amount of DNA originally present. In QPCR it is common to take the point where the fluorescence starts to exponentially increase i.e when one can tell signal from noise. Therefore a measurement of the initial amount of DNA in arbitrary units can be calculated as follows,

$$\hat{x} = e^{-c \ln 2},$$

where c is the number of cycles. To use this procedure on the mRNA harvested from the experiments the mRNA is converted to cDNA (complementary DNA).

A.3 Microarray experiments

There are numerous ways to measure gene expression including Northern blotting, differential display, serial analysis of gene expression and dot-blot analysis (van Hal *et al.*, 2000). The problem with all these techniques is that they are unsuitable for parallel testing the expression of multiple genes. The last ten years has seen the emergence of DNA "chips" or microarrays which enable the measurement of mRNA levels of thousands of genes simultaneously. Additional advantages of microarrays are that they are highly sensitive and small.

A.3.1 The microarray

A microarray consists of a reproducible pattern of thousands of different DNA strands attached to a solid support (Harrington *et al.*, 2000). The most popular type of microarray are Affymetrix GeneChips^{TM} which are synthesised high density oligonucleotide arrays. These microarrays are produced by synthesising tens of thousands of short oligonucleotides *in situ* onto glass wafers, one nucleotide at a time, using a modification of semiconductor photolithography technology (Macgregor and Squire, 2002). Each gene is

represented by between 15-20 different oligonucleotides on each chip (Harrington *et al.*, 2000). These oligonucleotides are specially designed to uniquely represent a gene. When the gene expression from all these oligonucleotides is combined an accurate measure of the expression is obtained. Unique to Affymetrix chips is that next to each oligonucleotide is a mismatch, an identical copy of the oligonucleotide except that its central nucleotide is changed to a different nucleotide. The amount of gene expression associated with the mismatch provides a measure of the background crosstalk that can occur for that particular oligonucleotide and hence this can be used to more accurately produce the exact expression signal for the gene.

Affymetrix produces a number of different microarrays each having a different composition of genes represented on the microarray. Here the Human U133A and B GeneChips^{TM} will be used. In combination these microarrays represent a total of 33,000 genes, which covers the vast majority of the human genome.

A.3.2 The experiment

The first step in the microarray experiment is to prepare the material that will be hybridised with the microarray. mRNA is extracted from the sample cells or tissues and the quality checked using standard techniques such as agarose gel electrophoresis or more recently though more accurate micro-capillary based devices such as the Agilent Bioanalyser (Macgregor and Squire, 2002). About 25-100 μg of RNA is required (Schulze and Downward, 2001) so normally the RNA is amplified. The mRNA is then converted to complimentary RNA (cRNA).

The next step is to label the RNA so that the quantity of RNA present can be measured. The most common approach is to attach a fluorescent label such as Cyanine dyes Cy3, Cy5 and to a lesser extent, fluorescein and rhodamine (van Hal *et al.*, 2000). After purification, the sample is hybridised to the microarrays at a suitable temperature. When the targets have been given long enough to bind to the microarray probes the microarrays are washed under stringent conditions to remove all non-specific binding. The microarrays are then read by a laser scanner to convert the fluorescence into a computer image. Each oligonucleotide and mismatch is then associated a value based on the average intensity in the areas that they are present.

A.3.3 Processing and quality control

To convert the oligonucleotide and mismatch values into a gene expression levels Affymetrix use the MAS5.0 algorithm (Affymetrix, 2002a,b). This processes the data to remove background intensities, normalises the data and converts the oligonucleotide levels into gene expression levels taking into account the cross-hybridisation that can occur.

After the expression levels have been obtained a series of quality control checks are made. First the image is visually examined for the presence of image artefacts for example

high/low intensity spots, scratches, or high regional/overall background. The checkerboard pattern around the microarray is also checked (this is produced by high levels of spiked in B2 Oligo). Additionally it is confirmed that the average background values and raw noise values are within a suitable range. To check that there is not degraded RNA or inefficient transcription of the cDNA the signals for the expression level from each end of the GAPDH and actin genes are compared. Another quality control check tests the sensitivity and linear range of the microarray by analysing the levels of biotin-labelled cRNA transcripts of bioB, bioC, bioD and cre are spiked into the microarray RNA sample in staggered quantities (1.5 pM, 5 pM, 25 pM and 100 pM respectively). The levels of these should be present and linearly correlated with their concentration.

A.3.4 Analysis

Microarray data produces a huge amount of data and as Schulze puts it "the challenge is to sieve through the mound of data to find meaningful results" (Schulze and Downward, 2001). This is not an easy task but many techniques have been produced to help. Perhaps the simplest and most effective method is to filter the data, which reduces the complication in mining the data by removing uninformative genes (Harrington *et al.*, 2000). The level and type of filtering depends on the type of experiment but it is common to filter out those genes with a change in normalised expression level below a particular threshold (Sebastiani and Ramoni, 2002) and those genes that have particularly high or low levels of expression. The majority of filtering is done in a very *ad hoc* manner with the thresholds set arbitrarily (Sebastiani and Ramoni, 2002).

One way to analyse the data is by trying to determine whether gene expression of a particular gene is significantly different in two experimental conditions, for example between cancerous and normal tissue. The most common approach is to use a t-test but there have been various advances over the years including "Significance Analysis of Microarrays" (SAM) (Tusher *et al.*, 2001) and various Bayesian methods (Baldi and Long, 2001; Efron *et al.*, 2001). By far the most common approach to analysing microarray data is to use cluster analysis (Schulze and Downward, 2001; Sherlock, 2000; Quackenbush, 2001; Sebastiani and Ramoni, 2002). Cluster analysis groups the data so that members of the same group are similar but between groups they are distinct. For gene expression data it is grouped in two ways:

1. By condition. In the case where there are multiple microarray experiments in different conditions, the experiments can be grouped according to the similarity of gene expression between experiments. An example is in cancer genomics where different cancer types can be grouped. This enables one to design suitable specific treatments for each group of cancer types based on their gene expression profiles.

2. By gene. In this situation the genes are grouped according to their individual expression profile across the experiments. The assumption behind this approach is

that genes with closely related expression patterns may be controlled by the same regulatory mechanisms (Schulze and Downward, 2001). If the genes show similar expression patterns over multiple different conditions one can conclude that they are functionally related, this is called 'guilt by association' (Schulze and Downward, 2001).

To be able to group data together one must be able to quantify the similarity between two sets of data and this is done by a distance measurement or metric. The most common ones are the Euclidean distance and Pearson correlation (Quackenbush, 2001). There are two types of cluster algorithm, those that require no outside input and group the data according to patterns in the data (unsupervised cluster analysis) and those that train the algorithm on a small subset of the data (training set) that is grouped by external information and then the rest of the data is sorted (supervised cluster analysis). Examples of unsupervised clustering are hierarchical clustering (Eisen *et al.*, 1998), k-means and self-organising maps (Kohonen, 1997). Examples of supervised clustering are naïve Bayes classifiers, support vector machines and neural networks.

A.3.5 K-means clustering

This method partitions the data into independent clusters using an iterative approach requiring hundreds of cycles before convergence. Like most other iterative approaches there is a danger that the clusters will not settle into the optimal configuration i.e. the solution gets trapped within a local minimum, so it is advisable to repeat the algorithm a number of times to ensure the global minimum is reached.

The process starts by the analyst choosing how many clusters the data should be divided into, k. k random vectors are created as 'cluster centroids', c_i. Each object is then assigned to its nearest cluster centroid. The cluster centroids are then recalculated using the mean or median of all the objects that belong to the cluster. The objects are then reassigned to their nearest centroid. The centroids and groupings are repeatedly calculated until some threshold in the change between iterations is met. To summarise, the whole technique is set on minimising the function,

$$J = \sum_{i=1}^{k} \sum_{j=1}^{n} u_{ij} d(x_j, c_i),$$

where u_{ij} is a matrix that takes values of 1 or 0 depending on whether object j is a member of cluster i. Other criteria functions J can be used, such as minimising the within-cluster variability whilst maximising the between-cluster variability (Sebastiani and Ramoni, 2002). In this method, the within-cluster variability is measured by the average distance between the cluster components and the cluster centroid, and between-cluster variability is measured by the average distance between the cluster components and all the centroids apart from their own cluster centroid.

Appendix B

Mathematical techniques

B.1 Least squares and maximum likelihood

In a statistical framework the notion of whether the parameters produce a good or bad fit is described as the probability that the data occurs given the parameters (and of course the model) (Press *et al.*, 2002). This is termed the *likelihood.* To get the best fit, the parameters are picked to maximise the likelihood; a maximum likelihood estimation. It is assumed that each data point, D_i, has a measurement error that is independently random and distributed as a Gaussian distribution and that the variance of the Gaussian is constant for all data points. The mean of the Gaussian distribution is assumed to be the point given by the parameter set θ and the variance is given by σ^2. Therefore the probability for one data point, D_i, given the model point, $x(\theta)_i$, is

$$\text{likelihood}_i = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{D_i - x(\theta)_i}{\sigma}\right]^2\right)$$

The joint probability density of all the data points is therefore

$$\text{likelihood} = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{D_i - x(\theta)_i}{\sigma}\right]^2\right)$$

and the log likelihood, $l(D, \theta)$

$$l(D, \theta) = -n \ln \sigma - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n} [D_i - x(\theta)_i]^2$$

Maximising the likelihood is equivalent to minimising the negative of the logarithm. As $l(D, \theta) \propto -\sum_{i=1}^{n} [D_i - x(\theta)_i]^2$ it can be concluded that maximising the likelihood (the maximum likelihood estimator) is equivalent to minimising the sum of the square differences (the least squares estimator).

B.2 Nelder-Mead downhill simplex optimisation method

A simplex is a geometric object that has one more vertex than there are dimensions, for example in a two dimensional space a simplex would be a triangle and in three dimensional space it would be a tetrahedron. This method takes an initial simplex and evaluates the function to be minimised at each vertex (Nelder and Mead, 1965). The highest value vertex is then moved according to various possible transformations (see Figure B.1) so that an optimal improvement is made in the vertex's function value. This process is repeated with the simplex tumbling downhill until it shrinks down at a minimum. This minimum will not necessarily be the global minimum. This method transforms the simplex to adapt it optimally to the local landscape so that the best steps can be made towards a minimum, for example in a long narrow valley the simplex will become long and thin and make long steps along the valley floor. Due to this behaviour the simplex has been described as amoeba-like, oozing down the valley to the minimum (Press *et al.*, 2002).

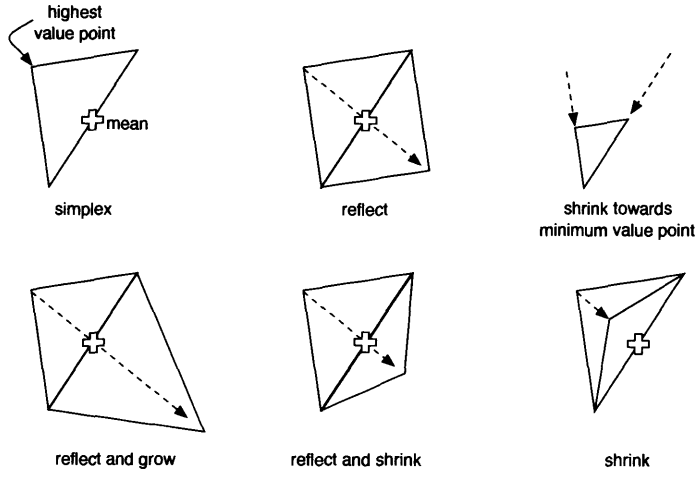


Figure B.1: This figure shows the various transformations that are performed in the downhill simplex method if the space was two dimensional. The top left diagram shows the simplex at the start of an iteration and the rest of the diagrams shows the possible state of the simplex after the iteration. Adapted from "The Nature of Mathematical Modelling", Gershenfeld (1999).

Consider an n dimensional space with a simplex whose points are $\mathbf{P}_0, \mathbf{P}_1, \ldots \mathbf{P}_n$. The function to be minimised is $f(\mathbf{P})$ and y_i is defined as the functional value at $\mathbf{P}_i$ i.e. $y_i = f(\mathbf{P}_i)$. The suffix l describes the point with the lowest function value, the suffix h indicates the point with the highest functional value and nh the point with the second highest function value. $\bar{\mathbf{P}}$ is the centroid of the all the simplex points apart from the point with the highest function value. All of the possible transformations used are described as follows:

$$\text{Transform}(\mathbf{P}_i, \mu) = \mu \mathbf{P}_i + (1 - \mu)\bar{\mathbf{P}}$$

where $\mathbf{P}_i$ is the point to be transformed. Depending on the transformation, μ can only take certain values.

Initially, a candidate point, $\mathbf{P}_t$, is proposed which is the reflection of $\mathbf{P}_h$ ($\mu = \alpha$ where $-1 < \alpha < 0$). If y_t is less than y_l a reflection and expansion is attempted on $\mathbf{P}_h$ giving $\mathbf{P}_{tt}$ ($\mu = \gamma$ where $\gamma < -1$). If $y_{tt} < y_l$ then point $\mathbf{P}_h$ is replaced by $\mathbf{P}_{tt}$, otherwise it is replaced by $\mathbf{P}_t$. If y_t is greater than y_l but less than y_{nh} then $\mathbf{P}_h$ is replaced by $\mathbf{P}_t$ and the iteration is exited. Otherwise, If y_t is less than y_h then $\mathbf{P}_h$ is replaced by $\mathbf{P}_t$ and the iteration is carried on.

If y_t is greater than y_l and greater than y_{nh} then a contraction is performed on $\mathbf{P}_h$ ($\mu = \beta$ where $0 < \beta < 1$) to give point $\mathbf{P}_{ttt}$. If y_{ttt} is less than y_h then $\mathbf{P}_h$ is replaced with $\mathbf{P}_{ttt}$, otherwise all the transformations have failed to produce a lowered function value for $\mathbf{P}_h$. In this situation the entire simplex is shrunk by half around the lowest point i.e. $\mathbf{P}_i = (\mathbf{P}_i + \mathbf{P}_l)/2$.

This procedure is repeated until the simplex meets some stopping criterion. Different values of α, β and γ are optimal depending on the function that needs to be minimised but Nelder and Mead (1965) suggest that the best general approach is to take the values $\alpha = -1$, $\beta = 0.5$ and $\gamma = -2$. These are also the values used by Press *et al.* (2002) and Gershenfeld (1999), and will be used here.

B.2.1 Testing the downhill simplex algorithm

It is important to ensure that the implementation of the Nelder-Mead algorithm is working correctly by running it on various test problems. The following minimisation problems were tested (these are the same as used by Nelder and Mead (1965)):

1. Rosenbrock's parabolic valley (Rosenbrock, 1960)

$$y = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$

   with starting position (-1.2,1) and minimum (1,1).

2. Powell's quartic function (Powell, 1962)

$$y = (x_1 + 10x_2)^2 + 5(x_3 - x_4)^2 + (x_2 - 2x_3)^4 + 10(x_1 - x_4)^4$$

   with starting position (3,-1,0,1) and minimum (0,0,0,0).

3. Fletcher and Powell's helical valley (Fletcher and Powell, 1963)

$$y = 100[x_3 - 10\theta(x_1, x_2)]^2 + \left[\sqrt{x_1^2 + x_2^2} - 1\right]^2 + x_3^2$$

   with starting position (-1,0,0) and minimum (-1,0,0).

The initial simplex is constructed from an initial point by taking a fixed length from the point along each dimension of the parameter space. The length scale λ was set to 3 and the accuracy to 10^{-5}. The downhill simplex method worked as expected (Table B.1). The number of steps it took to converge are different to those found in Nelder and Mead (1965) but are of the same order, this is due to a slightly different implementation.

Table B.1: Testing the downhill simplex routine on three test functions. The initial simplex was constructed by taking a fixed length, λ, along each dimension away from the starting point. λ =3 and the accuracy was set at 10^{-5}.

Test Number	1	2	3
Number of steps	82	251	142
Minimum value	2.3×10^{-16}	2.1×10^{-16}	0
Minimum point $(x_1 \ldots x_N)$	(0.999999987, 0.999999972)	$(-8.3 \times 10^{-5}, 8.3 \times 10^{-6}, -4.7 \times 10^{-5}, -4.7 \times 10^{-5})$	(-1,0,0)

B.3 Direction set (Powell's) method

A simple approach to multi-dimensional optimisation problems is to perform a series of line minimisations along a set of directions. For a parameter space with N dimensions the number of directions required in the direction set is N. Starting at an initial point $\mathbf{P}_0$, a line minimisation is performed along the first direction in the direction set. The system is then moved to $\mathbf{P}_1$, the minimum point given by the line minimisation. This is repeated along each direction in the direction set. A line minimisation along the first direction is then started again from point $\mathbf{P}_N$ and this process continues until a local minimum is found. The major implementation issue is choosing the direction set. A simple approach would be to choose the axes of the parameter space, but this approach can be very inefficient (Press *et al.*, 2002; Gershenfeld, 1999). Consider a two dimensional minimisation problem where there is a long thin valley that slopes to the minimum but is at an angle to the axes. The only way to get down this valley is to walk in a zig-zag pattern down the valley (Press *et al.*, 2002).

A more intelligent approach is used in Powell's method. This method updates the set of directions after each cycle through the set. To begin with the set of directions $\mathbf{u}_i$ are set to the axis directions. Then the direction set is cycled through as described above giving the point, $\mathbf{P}_N$. The direction in which the point has moved over the cycle is added to the direction set i.e. $\Delta\mathbf{P} = \mathbf{P}_N - \mathbf{P}_0$, and the direction that caused the *largest* drop in the function value is discarded. This may seem counter-intuitive but the largest drop direction is likely to be the major component of the new direction and so removing this direction keeps the directions as independent of each other as possible. A line minimisation is then performed along the direction $\Delta\mathbf{P}$ giving a point $\mathbf{P}_{N+1}$. The direction set is rearranged so that $\mathbf{u}_N = \Delta\mathbf{P}$. $\mathbf{P}_0$ is set to $\mathbf{P}_{N+1}$ and the process is

repeated until a minimum is achieved. The direction set is not altered in the following situations (Press *et al.*, 2002):

1. An extrapolation along $\Delta\mathbf{P}$ does not give an improved function value i.e. the downward slope in that direction has ended.
2. The decrease along $\Delta\mathbf{P}$ was not primarily due to any particular direction's decrease.
3. There is a substantial second derivative along $\Delta\mathbf{P}$ and it seems that the current point is near the minimum.

If this is the case then $\mathbf{P}_0$ is set to $\mathbf{P}_N$ and the process is restarted. Any line minimisation method could be used but here the Brent line minimisation method will be implemented (Press *et al.*, 2002).

B.4 Simulated annealing

A liquid has molecules that move freely with respect to each over. When the liquid is slowly cooled, this thermal mobility is slowly decreased and the atoms rearrange themselves to form a highly ordered structure called a crystal. This process is called annealing. This crystal is the minimum energy state possible for the system. If the system is rapidly cooled the system gets trapped in an amorphous state which has a higher energy state than the crystal. Thus in effect when a liquid is slowly cooled nature is solving a complex global minimisation problem whereas rapid cooling can be associated with a minimisation method that only finds a local minimum. The reason that the rapid cooling does not reach the global minimum energy state is that there is an energy barrier between the amorphous state and the crystal state. Slow cooling though, gives ample time for the thermal mobility to be used to explore many local rearrangements picking up on the configuration with the lowest energy state (Press *et al.*, 2002; Gershenfeld, 1999). The essence of this natural process can be used for an effective global minimisation method.

The first mathematical use of annealing was introduced by Metropolis *et al.* (1953) who used it to add thermodynamical fluctuations to statistical mechanical systems. They used the Boltzmann factor which states that the probability of seeing a state with energy E. $P(E)$ is proportional to the following:

$$p(E) \propto e^{-\frac{E}{kT}}$$

where T is the temperature and k is Boltzmann's constant. Therefore, if there are two states, one with energy E_1 and the other with energy E_2, then the probability of changing from state 1 to state 2 is given by

$$p = e^{-\Delta E/kT}$$

where $\Delta T = E_2 - E_1$. When $E_2 < E_1$ then $p > 1$ which is not possible so p is set to 1 (Press *et al.*, 2002). If the energy of a proposed state is less than the current state then the system will always move to that state but if the proposed state has a greater energy associated with it, it will only move with probability p. If the temperature is high then p will be high no matter what the energy difference, so virtually all states will be accepted. If, on the other hand, the temperature is small or zero the system will only move to those states with lower energy configurations. This replicates the desired effect of annealing.

Scott Kirkpatrick realised that the above formalisation could be used for other hard minimisation problems (Kirkpatrick *et al.*, 1983; Kirkpatrick, 1984) by replacing the energy of a state with some sort of cost measurement of a state. For example this cost measure could be the likelihood or the distance travelled (for example in the travelling salesman problem (Press *et al.*, 2002)). In this approach a state is proposed and the cost measure associated with this proposed state is measured. If the proposed state has a smaller cost measure than the current state then the system moves to the proposed state. If not, the proposed state is moved to with a probability $p = e^{-\Delta E/kT}$, where ΔE is the difference in cost measures and k is an arbitrary constant. This probabilistic decision is easily implemented computationally by taking a random number from a uniform distribution between zero and one. If the number is less than p the move is accepted, if it is greater than p it is rejected. The system is started with a high temperature so that virtually all proposed states are accepted whatever the cost measure. As the temperature is decreased the cost measure becomes more and more significant until the temperature reaches zero and the state moves to the nearest minimum. This method simulates the thermodynamic process of annealing and is called simulated annealing.

B.5 Adaptive step Runge-Kutta integrator

This integrator has the benefit of being able to adapt the step size so that when the solution is complex smaller step sizes are used, but when the solution is smooth larger step sizes are used. This means that the integrator will be as fast as possible whilst still retaining accuracy. To estimate the change required in the step size some measure of the truncation error, Δ_1, is required. This is calculated using the embedded Runge-Kutta formulae which were originally discovered by Fehlberg (1968). Fehlberg found both a fifth-order Runge-Kutta method and a fourth-order method that both required the same six function calls. The six function calls are as follows,

$$\begin{aligned} k_1 &= \delta t f(t_n, x_n), \\ k_2 &= \delta t f(t_n + a_2\delta t, x_n + b_{21}k_1), \\ &\cdots \\ k_6 &= \delta t f(t_n + a_6\delta t, x_n + b_{61}k_1 + \cdots + b_{65}k_5). \end{aligned}$$

This gives a fifth-order Runge-Kutta formula,

$$x_{t+1} = x_t + c_1k_1 + c_2k_2 + c_3k_3 + c_4k_4 + c_5k_5 + c_6k_6 + O(\delta t^6),$$

and a fourth-order embedded Runge-Kutta formula,

$$x^*_{t+1} = x_t + c^*_1k_1 + c^*_2k_2 + c^*_3k_3 + c^*_4k_4 + c^*_5k_5 + c^*_6.k_6 + O(\delta t^5).$$

Effective constant values have been determined by Cash and Karp (1990) and will be used here. The truncation error is then determined as follows,

$$\Delta_1 = x_{t+1} - x^*_{t+1}.$$

Normally x is actually a vector of values and so Δ_1 is also a vector of values. The scalar Δ_1 used in this situation is the component that has the largest value, the worst offender. The current timestep, δt_1, is then altered according to the truncation error and the desired accuracy, Δ_0,

$$\delta t_0 = \delta t_1 \left| \frac{\Delta_0}{\Delta_1} \right|^{0.2}.$$

The power of 0.2 is due to Δ_1 scaling as δt^5. If $\delta t_0 < \delta t_1$ then the step is retried with the new timestep, δt_0. When $\delta t_0 > \delta t_1$ the next step is processed with the new timestep. The desired accuracy can be defined in a number of ways but here it will be defined as,

$$\Delta_0 = \epsilon \times \left(|x_t| + \left| \delta t_1 \left. \frac{dx}{dt} \right|_{x_t,t} \right| \right),$$

where ϵ is some measure of the desired accuracy. The above is basically the fractional error except when x is very near zero.

There are a few adjustments that had to be made to make an effective implementation of the adaptive step size Runge-Kutta algorithm. Bounds have been placed on the step sizes, a minimum of 10^{-5} and a maximum of 5. Also a careful approach was implemented that dealt with the situation where the x values go out of bounds, in particular if the variables are moving to infinity the step is set to 10 and the variable is set to the upper bound.

Tests were performed comparing output from both the adaptive step routine and a more traditional Runge-Kutta routine to ensure that the adaptive step algorithm was working as expected.

Appendix C

Additional proofs and results

C.1 Proof that QR can be used to get the least squares result

For an over-defined set of linear equations it is impossible to solve exactly,

$$\mathbf{A} \cdot \mathbf{x} \approx \mathbf{b},$$

but we can find the "best" solution. The residual of this set of equations is defined as,

$$\mathbf{r} = \mathbf{A} \cdot \mathbf{x} - \mathbf{b}.$$

Here the least squares solution will be found i.e. the residual sum of squares (rss)(Dalziel, 1998).

$$\begin{aligned} rss &= \mathbf{r}^T \cdot \mathbf{r}, \\ &= [\mathbf{x}^T \cdot \mathbf{A}^T - \mathbf{b}^T][\mathbf{A} \cdot \mathbf{x} - \mathbf{b}], \\ &= \mathbf{x}^T \cdot \mathbf{A}^T \cdot \mathbf{A} \cdot \mathbf{x} - \mathbf{x}^T \cdot \mathbf{A}^T \cdot \mathbf{b} - \mathbf{b}^T \cdot \mathbf{A} \cdot \mathbf{x} + \mathbf{b}^T \cdot \mathbf{b}, \\ &= \mathbf{x}^T \cdot \mathbf{A}^T \cdot \mathbf{A} \cdot \mathbf{x} - 2\mathbf{x}^T \cdot \mathbf{A}^T \cdot \mathbf{b} + \mathbf{b}^T \cdot \mathbf{b}. \end{aligned}$$

To minimise rss the derivative is taken with respect to x and set to zero,

$$\frac{\partial rss}{\partial \mathbf{x}} = 2\mathbf{A}^T \cdot \mathbf{A} \cdot \mathbf{x} - 2\mathbf{A}^T \cdot \mathbf{b} = 0.$$

QR decomposition is now introduced for $\mathbf{A} = \mathbf{QR}$, using the fact that $\mathbf{Q}$ is orthogonal,

$$2\mathbf{A}^T \cdot \mathbf{Q} \cdot \mathbf{Q}^T \cdot \mathbf{A} \cdot \mathbf{x} - 2\mathbf{A}^T \cdot \mathbf{Q} \cdot \mathbf{Q}^T \cdot \mathbf{b} = 0,$$

$$2\mathbf{A}^T \cdot \mathbf{Q}[\cdot \mathbf{Q}^T \cdot \mathbf{A} \cdot \mathbf{x} - \mathbf{Q}^T \cdot \mathbf{b}] = 0.$$

Substituting for $\mathbf{R}$,

$$2\mathbf{A}^T \cdot \mathbf{Q}[\mathbf{R} \cdot \mathbf{x} - \mathbf{Q}^T \cdot \mathbf{b}] = 0,$$

For non-trivial solutions, the least squares is equivalent to solving,

$$\mathbf{R} \cdot \mathbf{x} = \mathbf{Q}^T \cdot \mathbf{b},$$

as required.

C.2 Quantifying error in the parameters when using QR decomposition

It is known that,

$$\sigma_\gamma^2 = \sum_{i=1}^{2n_d n_v} \sigma_i^2 \left(\frac{\partial \gamma}{\partial b_i}\right)^2, \tag{C.1}$$

for any parameter γ (Press *et al.*, 2002). When using QR decomposition to solve this problem, the following equation needs to be solved,

$$\mathbf{R} \cdot \mathbf{x} = \mathbf{Q}^T \cdot \mathbf{b},$$

which can be rearranged as follows,

$$\mathbf{x} = \mathbf{R}^{-1} \cdot \mathbf{Q}^T \cdot \mathbf{b}.$$

Writing this in element notation,

$$x_j = R_{jk}^{-1} Q_{ki}^T b_i.$$

Differentiating with respect to b_i gives,

$$\frac{\partial x_j}{\partial b_i} = R_{jk}^{-1} Q_{ki}^T.$$

Therefore substituting this into equation C.1 and displaying the sums explicitly gives,

$$\sigma^2(\gamma_j) = \sum_{i=1}^{2n_d n_v} \sigma_i^2 \left(\sum_{k=1}^{n_p+n_s} R_{jk}^{-1} Q_{ki}^T \right)^2.$$

C.3 Ensuring that the divergence is a real effect in Algorithm 3

To ensure that the effectiveness of Algorithm 2 and the divergent behaviour of Algorithm 3 were not model specific a new test model and data set were constructed. A

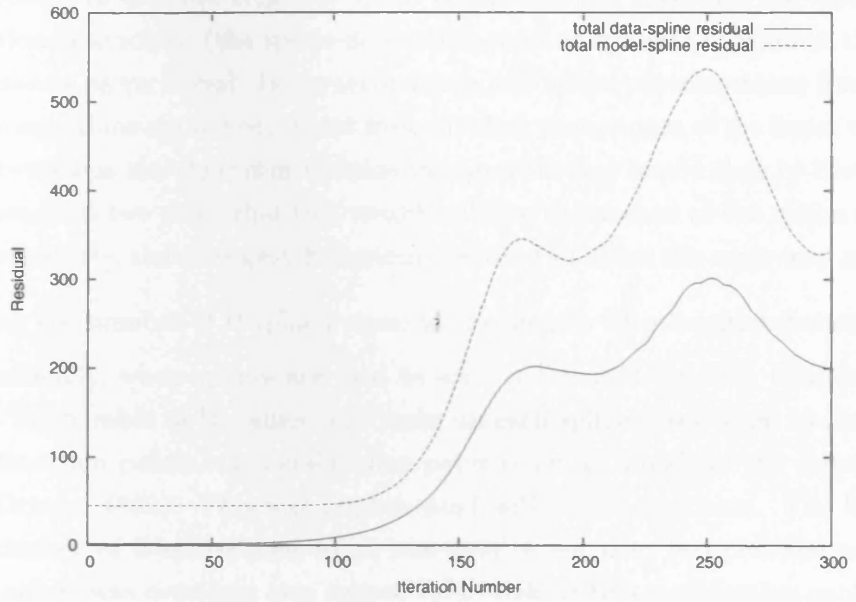


Figure C.1: A plot showing how the residuals evolve for the pendulum system when there is a total of 500 collocation points. 103 data points and 22 B-splines were used.

simple damped harmonic oscillator model was set up as follows,

$$\begin{aligned} \frac{dx}{dt} &= \gamma_1 y, \\ \frac{dy}{dt} &= -\gamma_2 x - \gamma_3 y, \end{aligned}$$

with trial parameters $\gamma_1 = 1$, $\gamma_2 = 1$ and $\gamma_3 = 0.1$. A data set with 103 time points was constructed between 0 and 21 time units and initial conditions $x = 10$, $y = 0$. It was found that model solution could be represented by 22 B-splines. With no error added to the data Algorithm 2 works well giving parameters of $\gamma_1 = 0.99994$, $\gamma_2 = 1.00003$ and $\gamma_3 = 0.10003$. Algorithm 2 also worked well when error was added producing reasonable values up to 100% error in the data, one example of the parameter estimates at 100% error are $\gamma_1 = 0.90614$, $\gamma_2 = 0.97947$ and $\gamma_3 = 0.05595$. When Algorithm 3 is applied to this model a similar type of behaviour is observed: when $n_c = 103$ the algorithm converges quickly in 17 iterations, but at $n_c = 500$ the residual diverges (Figure C.1). Therefore, the divergence effect is not model specific.

To examine further the possible cause of the divergence a few simple modifications to the algorithm were tested:

1. Solve the linear algebra problem in two steps.

   Instead of solving for all the parameters, both the model parameters and spline

coefficients, at once, this adaptation solves for one set of parameters and then the other. In the first step, the spline coefficients are fixed and the model-spline equations are solved (the spline-data is not used as it does not contain the model parameters as variables). In the second step, the model parameters are fixed and all of the equations are solved. Apart from dividing the solution of the linear equations into two steps the algorithm remains the same. It was hoped that by dividing the solution into two steps that this would stabilise the update of the spline estimate. Unfortunately, the divergent behaviour occurred in much the same way as before.

2. Setting the number of B-splines equal to the number of collocation points

   Traditionally, when splines are used to solve differential equation boundary problems, the number of B-splines that make up each spline is set equal to the number of collocation points, each collocation point is set at a node of the spline (Golub and Ortega, 1992). This was implemented within the algorithm. The limit that the number of B-splines had to be less than or equal to two plus the number of data points was overcome (see section 7.7). Making this modification could potentially help the convergence of the spline because at the B-spline nodes there is the greatest control over the position and gradient of the spline. By letting the spline-model equations be evaluated at this point it was hoped that the spline would be stabilised. The computer time for the algorithm is greatly increased because of the number of B-splines. Unfortunately this did not alter the behaviour of the system and the divergence persisted.

3. Using the spline produced by the unmodified method as the initial spline for the modified method.

   It was important to check that the divergence was not caused by the way that the initial spline was set up. To test this, the algorithm was changed so that the original algorithm was run first (using an estimate vector) and the solution spline produced would be used as the initial spline for the modified algorithm. It is known that the solution spline converges in this situation, so it is a good choice as the starting position. This change made little difference to the behaviour and the spline still diverged when there was a high number of collocation points. This suggests that the problem does not emerge from the way the initial spline is set up.

C.4 A method to detect whether Algorithm 3 is divergent

To test whether the spline has converged to a final state within some tolerance, the difference between the new and old spline coefficients is used,

$$\text{Spline Difference} = \sum_{j=0}^{n_s-1} \sum_{i=0}^{n_v-1} \left(b_{ijq} - b_{ij(q-1)}\right)^2, \tag{C.2}$$

where b_{ijq} is the ith component of $\mathbf{b}_{jq}$, which is the spline coefficients vector for the qth iteration. If this Spline Difference is less than 10^{-8} at any iteration then the spline is assumed to have converged to a steady state. This spline difference can also be used to test whether the algorithm is diverging; at regular intervals of 20 iterations the average of the last 5 spline differences are calculated, if this average is greater than the previously calculated average then the algorithm is defined as having diverged and the algorithm can stop. Even though the choice of interval and the number included in the average are arbitrarily chosen, in practice it was found to work well.

C.5 Proof of perfect correlation if two genes share the same transcription factor

Suppose that there are two genes that share the same transcription factor, there $G_g(t)$ values will be,

$$G_1(t) = B_1 + S_1 f(t),$$

$$G_2(t) = B_2 + S_2 f(t).$$

The correlation between G_1 and G_2 is defined as follows,

$$\langle G_1, G_2 \rangle = \frac{Cov(G_1, G_2)}{\sqrt{Var(G_1)Var(G_2)}},$$

where $Cov(G_1, G_2)$ is the covariance between G_1 and G_2 and $Var(G_1)$ is the variance of G_1. It is assumed that the shared transcription factor $f(t)$ has a mean of μ and a variance of σ^2. The variance can be defined as follows,

$$Var(x) = E(x^2) - E(x)^2,$$

where $E(x)$ is the expected value of x. From this it follows that,

$$E(f(t)^2) = \sigma^2 + \mu^2,$$

and that,

$$Var(G_1) = E(G_1^2) - (E(G_1))^2,$$

$$\Rightarrow Var(G_1) = E((B_1 + S_1 f(t))^2) - (E(B_1 + S_1 f(t))^2,$$

$$\Rightarrow Var(G_1) = E(B_1^2 + 2S_1 f(t) + S_1^2 f(t)^2) - (B_1 + S_1\mu)^2,$$

$$\Rightarrow Var(G_1) = B_1^2 + 2S_1\mu + S_1^2\sigma^2 + S_1^2\mu^2 - B_1^2 - 2S_1\mu - S_1^2\mu^2,$$

$$\Rightarrow Var(G_1) = S_1^2\sigma^2.$$

Similarly, $Var(G_2) = S_2^2\sigma^2$. The covariance is defined as follows,

$$Cov(G_1, G_2) = E(G_1G_2) - E(G_1)E(G_2)$$

$$\Rightarrow Cov(G_1, G_2) = E(B_1B_2 + B_2S_1f(t) + B_1S_2f(t) + S_1S_2f(t)^2) - E(G_1)E(G_2)$$

$$\Rightarrow Cov(G_1, G_2) = B_1B_2 + B_2S_1\mu + B_1S_2\mu + S_1S_2(\sigma^2 + \mu^2) - (B_1 + S_1\mu)(B_2 + S_2\mu)$$

$$\Rightarrow Cov(G_1, G_2) = S_1S_2\sigma^2.$$

Therefore, bringing this all together,

$$\langle G_1, G_2 \rangle = \frac{S_1S_2\sigma^2}{\sqrt{S_1^2\sigma^2 S_2^2\sigma^2}}$$

$$\Rightarrow \langle G_1, G_2 \rangle = 1.$$

If two genes share the same transcription factor and the model is obeyed the correlation between the two $G_g(t)$s will be perfect.

C.6 Estimated degradation rates from non-irradiated data

The experimental procedure in section 8.2.2 was used to gather degradation microarray time series data but this time the cells were not irradiated. Cells were harvested and measured at 0, 0.25, 0.5, 1, 1.5, 2, 2.5, 3, 4, 5, 6 and 8 hours. QPCR measurements were also made on the non-irradiated cells for a number of genes (Table C.1).

Table C.1: The degradation rate constants found from the QPCR data when the cells were not irradiated.

Gene	Degradation rate constant
CD71	0.232
HPRT1	0.256
PGK1	0.224

Table C.2: The total scaling factors used on the non-irradiated microarray degradation data.

Time	0	0.25	0.5	1	1.5	2	2.5	3	4	5	6	8
α	1	0.945	0.863	0.909	0.839	0.754	0.615	0.630	0.455	0.356	0.335	0.247

The data was re-normalised and as with the irradiated data the total scaling factor decreases as the time increases (Table C.2). Interestingly, it appears that less adjustment has to be made in the non-irradiated case, which could indicate that on average there is a slower rate of decline of mRNA. This would make sense if the initial amount of total mRNA is lower in the non-irradiated microarray degradation data and this is expected.

Table C.3: A table to show the degradation rates estimated from the microarray data before and after the adjustment taking into account QPCR data. This is for non-irradiated data.

Affymetrix label	Gene name	Degradation rates		
		QPCR	Microarray before	Microarray after
207332_s_at	CD71	0.232	0.175	0.281
202854_at	HPRT1	0.256	0.182	0.280
200737_at	PGK1	0.224	0.0970	0.207
200738_s_at	PGK1	0.224	0.106	0.215
217356_s_at	PGK1	0.224	0.0955	0.206
221616_s_at	PGK1	0.224	0.131	0.245

Table C.4: A table to show the number of genes (out of 22277) that have data that have "bad" fits to the model of degradation.

	Number of genes with p			
	$< 10^{-5}$	$< 10^{-4}$	< 0.001	< 0.01
irradiated data	182	333	719	1807
non-irradiated data	1053	1560	2429	3938

The data was then anchored to the estimated degradation rates found from the non-irradiated data QPCR data, the total adjustment to the degradation rate was 0.108 (Table C.3).

The degradation rates can be very different between the irradiated and non-irradiated data. For the QPCR results the non-irradiated degradation rates are at least half of the irradiated values (see Table 8.1 & Table C.3). The difference in the degradation rates are also shown in the adjusted microarray degradation rates. The average degradation rate across the whole of the microarray is 0.458 for the non-irradiated data compared to the irradiated's 0.674. The average ratio of the estimated irradiated rate to the non-irradiated rate is 2.234. These are major discrepancies that might be caused by the simplicity of the degradation model or something in the experiment design.

The degradation rates estimated from the irradiated data are considered to be of better quality than the non-irradiated estimates for the following reasons:

1. The non-irradiated data produces a greater amount of poor degradation model fits. This can be seen in the number of low p values (see Table C.4) and the average p for all the genes, which is higher for the irradiated data (0.539) than for the non-irradiated data (0.461).

2. The non-irradiated results have a considerably higher variance than the irradiated results, 0.217 to 0.070, an order of magnitude different.

3. The irradiated data has a greater number of negative degradation rate constants (220 to 18), which indicates that the data is noisy and the estimation is not accurate.

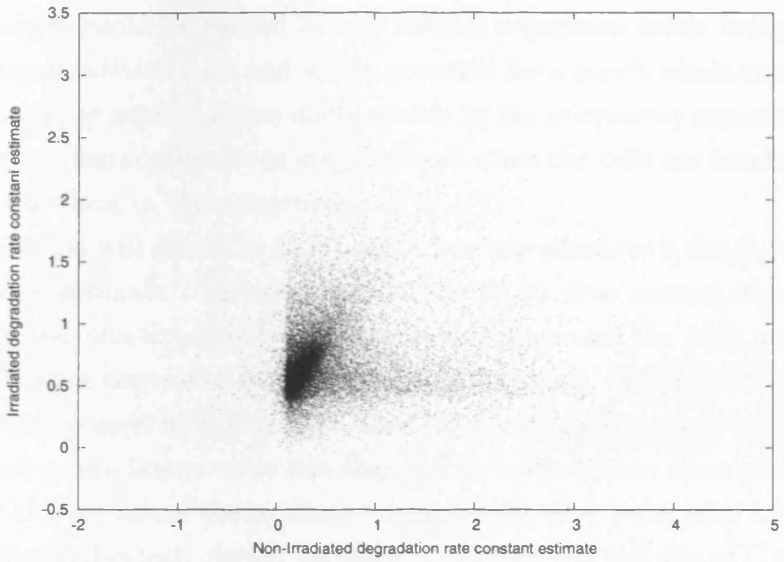


Figure C.2: A plot depicting the distribution of the degradation rates found for the irradiated and non-irradiated microarray data.

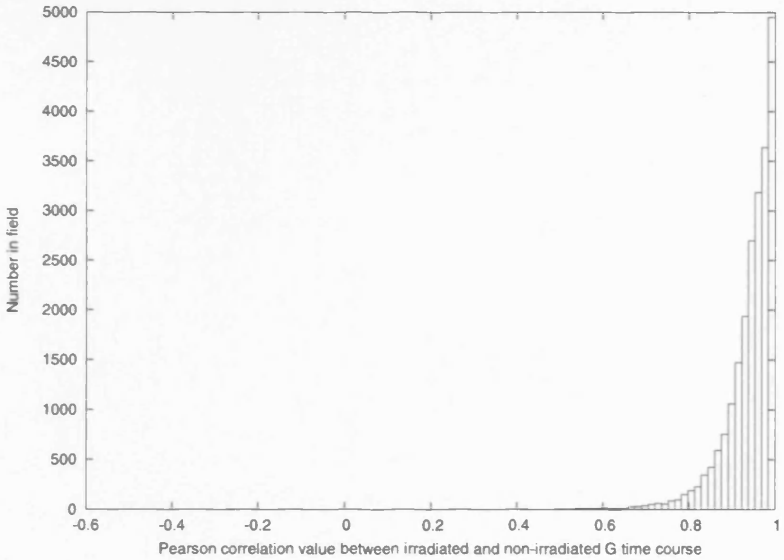


Figure C.3: A plot showing the distribution of correlation values between the G time courses with degradation rates estimated from irradiated data and non-irradiated data.

4. The non-irradiated estimates have a lot more high values (491 greater than 2) than the irradiated estimates (50 greater than 2) (see Figure C.2).

These poor effects could be caused by the mRNA expression levels being considerably lower in the non-irradiated data and so the potential for a gene's whole expression profile to be in the low noisy zone and thus undetectable by the microarray experiment is greatly increased. Better degradation rates are produced when the cells are irradiated and so it is sensible to use them in the construction of $G_g(t)$.

The degradation will affect the $G_g(t)$ value, but how sensitive is the $G_g(t)$ value to the degradation rate estimate? To examine this, the $G_g(t)$ time courses constructed using the irradiated and non-irradiated degradation estimates and the 5Gy microarray data (see chapter 3) were correlated for each of the 22283 genes. Generally the correlation is very high as can be seen in Figure C.3. Less than 0.1% had a negative correlation and only 3.57% had a correlation value less than 0.8 (it is likely that these poor correlations are caused by at least one of the estimated degradation rates being poor i.e. having a low p or a high estimated error). As the correlation between the two sets of G time series are good then it appears that the degradation rates estimated are satisfactory for the final analysis. This is probably because the time course $x_g(t)$ has a much greater effect on the $G_g(t)$ time course than the degradation rate. Equation 8.2 shows that $x_g(t)$ is used both in the estimate of the derivative *and* the degradation rate term.

Appendix D

Additional tables

D.1 Additional tables of results for simulated annealing

Table D.1: The results from using simulated annealing with downhill simplex parameter estimation. A range of initial temperatures, estimated counts and initial points were used.

Initial point	Initial temp	Estimated count (K)	No. of iterations	Least squares value	D_{ATM}	D_{MDM2}	D_{p53}	p_{MDM2}	p_{p53}	k_1	k_2	k_3	k_4
					0.05	0.18	0.041	0.01	0.52	1.42	0.39	2.50	0.75
(b)	10	10^6	1000000	6.36×10^{-4}	0.0500	0.200	0.0445	0.0187	0.521	1.42	0.391	2.53	0.755
(b)	10	10^6	1000000	6.36×10^{-4}	0.0500	0.200	0.0445	0.0187	0.521	1.42	0.391	2.53	0.755
(b)	10	10^6	1000000	6.36×10^{-4}	0.0500	0.200	0.0445	0.0187	0.521	1.42	0.391	2.53	0.755
(b)	10	10^6	1000000	6.36×10^{-4}	0.0500	0.200	0.0445	0.0187	0.521	1.42	0.391	2.53	0.755
(b)	10	10^6	1000000	0.288	0.0446	0.215	0.0617	0.0119	0.522	1.39	0.376	2.66	0.698
(b)	10	10^7	10000000	6.36×10^{-4}	0.0500	0.200	0.0445	0.0187	0.521	1.42	0.391	2.53	0.755
(b)	10	10^7	10000000	0.237	0.0553	0.191	0.0260	0.0257	0.519	1.45	0.406	2.42	0.800
(b)	100	10^6	1000000	6.36×10^{-4}	0.0500	0.200	0.0445	0.0187	0.521	1.42	0.391	2.53	0.755
(b)	100	10^6	1000000	6.11	0.0505	-39.6	835	-289	589	107	417	10600	2780
(b)	100	10^6	1000002	16.4	1810	1110	1100	469	359	-1300	-208	-340	-561
(b)	100	10^7	10000000	11.4	1870	652	1390	157	758	232	-141	1520	1870
(b)	1000	10^6	1000000	110	1640	517	1520	115	565	-852	436	-8010	34800
(c)	10	10^6	1000001	5.22	0.0534	626	1160	313	307	-1300	235	1820	854
(c)	10	10^6	1000001	5.39	0.0495	1540	1590	666	806	-690	559	0.0208	13.7
(c)	10	10^6	1000000	5.48	0.0484	945	2500	597	604	-2970	418	3.28	696
(c)	10	10^6	1000000	5.77	0.0455	1510	-1080	520	267	2990	187	4200	176
(c)	10	10^6	1000000	5.98	0.0502	1490	1240	619	567	-666	526	-1340	-197
(c)	10	10^7	10000000	5.37	0.0501	2160	112	3500	56.1	-28.7	43.8	-29200	-2160
(d)	10	10^6	1000000	4.28	0.0519	1870	-333	726	143	1650	192	-2100	-1100
(d)	10	10^6	1000000	5.50	0.0497	1650	1490	360	960	-177	579	191	145
(d)	10	10^6	1000001	5.67	0.0505	970	1830	595	526	-2000	403	-413	547
(d)	10	10^6	1000002	5.71	0.0484	1720	1750	1200	296	-2311	247	-4660	-360
(d)	10	10^6	1000000	12.4	1840	-44	-1973	-23.9	136	4660	-848	-143	314

Table D.2: The results from using simulated annealing with downhill simplex parameter estimation when estimating the model solution with B-splines. A range of initial temperatures and estimated counts used. (b) was used as the initial point.

Starting temp	Estimated count (K)	Number of iterations	Least squares value	D_{ATM}	D_{MDM2}	D_{p53}	p_{MDM2}	p_{p53}	k_1	k_2	k_3	k_4
				0.05	0.18	0.041	0.01	0.52	1.42	0.39	2.50	0.75
10	10^7	10353048	0.0289	0.0466	0.273	0.485	0.179	0.510	0.550	0.369	1.12	0.800
10	10^7	10323240	0.0165	0.0515	0.452	0.274	0.213	0.492	0.892	0.371	1.37	0.558
10	10^7	10413608	0.0411	0.0438	0.123	0.257	0.0847	0.412	0.680	0.297	0.270	0.287
10	10^7	10302822	0.0240	0.0535	0.111	0.766	0.0184	0.585	0.245	0.422	1.35	0.448
10	10^7	10349138	0.0215	0.0507	0.216	0.143	0.109	0.442	0.994	0.335	0.793	0.392
10	10^8	100092291	0.00448	0.0519	0.161	0.0205	0.0100	0.501	1.40	0.378	1.68	0.400
10	10^8	100207829	0.00223	0.0490	0.151	0.184	0.0190	0.512	1.13	0.382	2.04	0.67
10	10^8	100207386	0.00202	0.0498	0.115	0.180	4.66×10^{-5}	0.504	1.11	0.373	2.13	0.716
10	10^9	10^9	0.0193	0.0475	0.146	0.522	0.0777	0.578	0.695	0.419	1.34	0.694
100	10^7	10608630	0.122	0.0943	0.657	0.731	0.327	0.887	1.16	0.779	1.60	0.774
100	10^7	10431066	0.0648	0.106	0.0465	0.134	0.0907	0.310	0.572	0.304	0.854	0.841
100	10^7	10379737	0.0606	0.0819	0.605	0.599	0.313	0.505	0.358	0.293	0.723	0.373
100	10^7	10309775	0.144	0.0283	1.05	1.03	0.595	0.872	0.506	0.824	0.940	0.785
100	10^7	10376820	0.123	0.0935	0.448	0.698	0.368	0.873	1.21	0.723	0.431	1.07

D.2 Table of the top 150 predicted p53 targets

Table D.3: The top 150 predicted p53 targets.

Affymetrix code	Description	Gene Tag	Correlation	Verification Score
218346_s_at	sestrin 1	SESN1	0.963	3.90
205780_at	BCL2-interacting killer	BIK	0.960	6.57
203409_at	damage-specific DNA binding protein 2, 48kDa	DDB2	0.954	10.7
209295_at	tumour necrosis factor receptor superfamily, member 10b	TNFRSF10B	0.921	6.52
218627_at	hypothetical protein FLJ11259	FLJ11259	0.896	3.56
202284_s_at	cyclin-dependent kinase inhibitor 1A (p21, Cip1)	CDKN1A	0.893	8.07
201834_at	protein kinase, AMP-activated, beta 1 non-catalytic subunit	PRKAB1	0.888	6.30
204674_at	lymphoid-restricted membrane protein	LRMP	0.865	3.40
218403_at	p53-inducible cell-survival factor	P53CSV	0.852	7.75
212371_at	CGI-146 protein	PNAS-4	0.850	2.61
213293_s_at	tripartite motif-containing 22	TRIM22	0.847	6.07
208796_s_at	cyclin G1	CCNG1	0.844	5.18
215719_x_at	Fas (TNF receptor superfamily, member 6)	FAS	0.830	8.11
219628_at	p53 target zinc finger protein	WIG1	0.819	3.70
212815_at	activating signal cointegrator 1 complex subunit 3	ASCC3	0.817	5.93
216252_x_at	Fas (TNF receptor superfamily, member 6)	FAS	0.814	4.54
205692_s_at	CD38 antigen (p45)	CD38	0.798	9.02
203725_at	growth arrest and DNA-damage-inducible, alpha	GADD45A	0.798	11.0
204566_at	protein phosphatase 1D magnesium-dependent, delta isoform	PPM1D	0.798	6.05
212430_at	RNA-binding region (RNP1, RRM) containing 1	RNPC1	0.771	2.33
213038_at	IBR domain containing 3	IBRDC3	0.769	2.79
219361_s_at	hypothetical protein FLJ12484	FLJ12484	0.755	5.43
218751_s_at	F-box and WD-40 domain protein 7	FBXW7	0.746	-0.541
207813_s_at	ferredoxin reductase	FDXR	0.745	7.72
201835_s_at	protein kinase, AMP-activated, beta 1 non-catalytic subunit	PRKAB1	0.744	5.92
202726_at	ligase I, DNA, ATP-dependent	LIG1	0.739	2.69
218007_s_at	ribosomal protein S27-like	RPS27L	0.736	9.36
218031_s_at	checkpoint suppressor 1	CHES1	0.734	1.15
207426_s_at	tumour necrosis factor (ligand) superfamily, member 4	TNFSF4	0.723	5.26
202181_at	KIAA0247	KIAA0247	0.720	2.22
203562_at	fasciculation and elongation protein zeta 1 (zygin I)	FEZ1	0.718	-1.86
218527_at	aprataxin	APTX	0.707	-2.32
209375_at	xeroderma pigmentosum, complementation group C	XPC	0.703	5.80
207616_s_at	TRAF family member-associated NFKB activator	TANK	0.703	-0.617
200730_s_at	protein tyrosine phosphatase type IVA, member 1	PTP4A1	0.700	4.45
211318_s_at	RAE1 RNA export 1 homolog (S. pombe)	RAE1	0.695	3.44
217743_s_at	transmembrane protein 30A	TMEM30A	0.692	2.33
204780_s_at	Fas (TNF receptor superfamily, member 6)	FAS	0.691	7.78
205349_at	guanine nucleotide binding protein (G protein), alpha 15	GNA15	0.689	5.15
201093_x_at	succinate dehydrogenase complex, subunit A, flavoprotein	SDHA	0.686	0.716
218288_s_at	hypothetical protein MDS025	MDS025	0.684	2.38
214771_x_at	myosin phosphatase-Rho interacting protein	M-RIP	0.679	-0.935
203846_at	tripartite motif-containing 32	TRIM32	0.676	0.862
35974_at	lymphoid-restricted membrane protein	LRMP	0.673	3.69
36564_at	IBR domain containing 3	IBRDC3	0.673	3.19
219815_at	galactose-3-O-sulfotransferase 4	GAL3ST4	0.671	3.12
205531_s_at	glutaminase 2 (liver, mitochondrial)	GLS2	0.663	2.52
219099_at	chromosome 12 open reading frame 5	C12orf5	0.660	6.49
204060_s_at	protein kinase, X-linked /// protein kinase, Y-linked	PRKX	0.653	2.28
219627_at	hypothetical protein FLJ12700	FLJ12700	0.652	0.801
207760_s_at	nuclear receptor co-repressor 2	NCOR2	0.651	-0.635
203509_at	sortilin-related receptor, L(DLR class) A repeats-containing	SORL1	0.651	1.70
202695_s_at	serine/threonine kinase 17a (apoptosis-inducing)	STK17A	0.651	6.50
218634_at	pleckstrin homology-like domain, family A, member 3	PHLDA3	0.649	3.71
201012_at	annexin A1	ANXA1	0.648	1.90
209178_at	DEAH (Asp-Glu-Ala-His) box polypeptide 38	DHX38	0.647	-3.14
203573_s_at	Rab geranylgeranyltransferase, alpha subunit	RABGGTA	0.646	0.285
216026_s_at	polymerase (DNA directed), epsilon	POLE	0.645	-4.03
213030_s_at	plexin A2	PLXNA2	0.644	4.23
200921_s_at	B-cell translocation gene 1, anti-proliferative	BTG1	0.644	5.42
201236_s_at	BTG family, member 2	BTG2	0.644	2.50
206066_s_at	RAD51 homolog C (S. cerevisiae)	RAD51C	0.643	2.61
201639_s_at	cleavage and polyadenylation specific factor 1, 160kDa	CPSF1	0.643	-1.82
202481_at	dehydrogenase/reductase (SDR family) member 3	DHRS3	0.640	1.56
210705_s_at	tripartite motif-containing 5	TRIM5	0.639	1.45
204573_at	carnitine O-octanoyltransferase	CROT	0.639	3.80
203578_s_at	solute carrier family 7, member 6	SLC7A6	0.636	1.05
220623_s_at	testis specific, 10	TSGA10	0.636	0.472

Continued on Next Page...

Table D.3 – Continued

Affymetrix code	Description	Gene Tag	Correlation	Verification Score
218014_at	pericentrin 1	PCNT1	0.633	0.294
214760_at	zinc finger protein 337	ZNF337	0.633	0.305
209626_s_at	oxysterol binding protein-like 3	OSBPL3	0.631	1.91
37860_at	zinc finger protein 337	ZNF337	0.630	-0.970
204526_s_at	TBC1 domain family, member 8	TBC1D8	0.630	0.662
215411_s_at	TRAF3 interacting protein 2	TRAF3IP2	0.628	6.64
208642_s_at	X-ray repair complementing defective repair in Chinese hamster cells 5	XRCC5	0.627	0.651
212607_at	v-akt murine thymoma viral oncogene homolog 3	AKT3	0.626	1.53
201717_at	mitochondrial ribosomal protein L49	MRPL49	0.624	1.24
214950_at	interleukin 9 receptor	IL9R	0.621	-1.47
201714_at	tubulin, gamma 1	TUBG1	0.620	0.0246
210349_at	calcium/calmodulin-dependent protein kinase IV	CAMK4	0.618	0.952
209917_s_at	TP53 activated protein 1	TP53AP1	0.612	4.05
218167_at	hypothetical protein LOC51321	LOC51321	0.611	1.22
219284_at	HSPB associated protein 1	HSPBAP1	0.611	-0.764
202693_s_at	serine/threonine kinase 17a	STK17A	0.610	8.63
206571_s_at	mitogen-activated protein kinase kinase kinase kinase 4	MAP4K4	0.609	1.88
201186_at	low density lipoprotein receptor-related protein associated protein 1	LRPAP1	0.609	0.664
218902_at	Notch homolog 1, translocation-associated	NOTCH1	0.608	1.633
203518_at	lysosomal trafficking regulator	LYST	0.608	-1.00
221081_s_at	hypothetical protein FLJ22457	FLJ22457	0.608	6.33
205266_at	leukemia inhibitory factor	LIF	0.607	3.42
201700_at	cyclin D3	CCND3	0.607	-0.0683
200041_s_at	HLA-B associated transcript 1	BAT1	0.606	-1.86
209147_s_at	phosphatidic acid phosphatase type 2A	PPAP2A	0.602	-0.425
217804_s_at	interleukin enhancer binding factor 3, 90kDa	ILF3	0.595	2.07
200732_s_at	protein tyrosine phosphatase type IVA, member 1	PTP4A1	0.592	2.87
208634_s_at	microtubule-actin crosslinking factor 1	MACF1	0.592	0.522
204781_s_at	Fas (TNF receptor superfamily, member 6)	FAS	0.585	6.42
201939_at	polo-like kinase 2 (Drosophila)	PLK2/SNK	0.584	3.62
204683_at	intercellular adhesion molecule 2	ICAM2	0.582	1.28
204391_x_at	transcriptional intermediary factor 1	TIF1	0.580	1.21
213479_at	neuronal pentraxin II	NPTX2	0.579	-1.64
212793_at	dishevelled associated activator of morphogenesis 2	DAAM2	0.576	-0.199
217716_s_at	Sec61 alpha 1 subunit (S. cerevisiae)	SEC61A1	0.575	0.884
210886_x_at	TP53 activated protein 1	TP53AP1	0.575	2.88
209051_s_at	ral guanine nucleotide dissociation stimulator	RALGDS	0.571	-0.637
212754_s_at	KIAA1040 protein	KIAA1040	0.571	-1.22
204205_at	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3G	APOBEC3G	0.570	1.09
202585_s_at	nuclear transcription factor, X-box binding 1	NFX1	0.568	1.49
211974_x_at	recombining binding protein suppressor of hairless	RBPSUH	0.567	1.00
201301_s_at	annexin A4	ANXA4	0.566	1.09
33132_at	cleavage and polyadenylation specific factor 1, 160kDa	CPSF1	0.564	-1.44
203412_at	leucine-zipper-like transcription regulator 1	LZTR1	0.563	-2.39
203946_s_at	arginase, type II	ARG2	0.562	0.556
219347_at	nudix (nucleoside diphosphate linked moiety X)-type motif 15	NUDT15	0.562	0.690
202702_at	tripartite motif-containing 26	TRIM26	0.560	1.13
202281_at	cyclin G associated kinase	GAK	0.558	-0.506
212975_at	KIAA0870 protein	KIAA0870	0.557	0.772
213829_x_at	tumour necrosis factor receptor superfamily, member 6b	TNFRSF6B	0.556	0.304
206060_s_at	protein tyrosine phosphatase, non-receptor type 22 (lymphoid)	PTPN22	0.556	2.78
209498_at	carcinoembryonic antigen-related cell adhesion molecule 1	CEACAM1	0.555	2.63
218562_s_at	hypothetical protein FLJ10747	FLJ10747	0.553	0.494
208066_s_at	general transcription factor IIB	GTF2B	0.553	2.12
202822_at	LIM domain containing preferred translocation partner in lipoma	LPP	0.553	-2.10
204178_s_at	RNA binding motif protein 14	RBM14	0.553	1.09
200000_s_at	PRP8 pre-mRNA processing factor 8 homolog (yeast)	PRPF8	0.550	-1.30
211012_s_at	promyelocytic leukemia	PML	0.549	-1.52
200802_at	seryl-tRNA synthetase	SARS	0.549	2.50
212862_at	CDP-diacylglycerol synthase 2	CDS2	0.548	3.48
205854_at	tubby like protein 3	TULP3	0.548	-0.0668
203579_s_at	solute carrier family 7, member 6	SLC7A6	0.547	2.56
209837_at	adaptor-related protein complex 4, mu 1 subunit	AP4M1	0.546	-0.412
217732_s_at	integral membrane protein 2B	ITM2B	0.546	-0.192
215855_s_at	TATA element modulatory factor 1	TMF1	0.545	-0.282
222191_s_at	xylosylprotein beta 1,4-galactosyltransferase, polypeptide 7	B4GALT7	0.543	-0.665
218168_s_at	chaperone, ABC1 activity of bc1 complex like (S. pombe)	CABC1	0.542	0.548
202395_at	N-ethylmaleimide-sensitive factor	NSF	0.541	2.00
204620_s_at	chondroitin sulfate proteoglycan 2 (versican)	CSPG2	0.540	-1.28
201390_s_at	casein kinase 2, beta polypeptide	CSNK2B	0.539	2.94

Continued on Next Page...

Table D.3 - Continued

Affymetrix code	Description	Gene Tag	Correlation	Verification Score
211630_s_at	glutathione synthetase	GSS	0.539	1.06
53987_at	RAN binding protein 10	RANBP10	0.539	-0.944
212541_at	FAD-synthetase	PP591	0.538	-1.43
213541_s_at	v-ets erythroblastosis virus E26 oncogene like (avian)	ERG	0.537	0.628
212788_x_at	ferritin, light polypeptide	FTL	0.537	2.47
207541_s_at	exosome component 10	EXOSC10	0.536	-1.51
204061_at	protein kinase, X-linked	PRKX	0.535	2.17
201952_at	—	—	0.533	-0.573
219863_at	hect domain and RLD 5	HERC5	0.533	1.91
38269_at	protein kinase D2	PRKD2	0.532	1.25
218183_at	chromosome 16 open reading frame 5	C16orf5	0.532	1.09
204215_at	chromosome 7 open reading frame 23	C7orf23	0.532	0.373

D.3 Ranked lists for the predicted targets of the training sets

Table D.4: Ranked list of targets for training set 1.

Affymetrix tag	Description	Gene Symbol	Correlation
202644_s_at	tumour necrosis factor, alpha-induced protein 3	TNFAIP3	0.982
202643_s_at	tumour necrosis factor, alpha-induced protein 3	TNFAIP3	0.977
209795_at	CD69 antigen (p60, early T-cell activation antigen)	CD69	0.976
201502_s_at	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha	NFKBIA	0.970
202687_s_at	tumour necrosis factor (ligand) superfamily, member 10	TNFSF10	0.961
202688_at	tumour necrosis factor (ligand) superfamily, member 10	TNFSF10	0.961
216620_s_at	Rho guanine nucleotide exchange factor (GEF) 10	ARHGEF10	0.955
204440_at	CD83 antigen (activated B lymphocytes, immunoglobulin superfamily)	CD83	0.946
203752_s_at	jun D proto-oncogene	JUND	0.925
201464_x_at	v-jun sarcoma virus 17 oncogene homolog (avian)	JUN	0.922
204702_s_at	nuclear factor (erythroid-derived 2)-like 3	NFE2L3	0.913
205463_s_at	platelet-derived growth factor alpha polypeptide	PDGFA	0.911
214722_at	Notch homolog 2 (Drosophila) N-terminal like	NOTCH2NL	0.904
217850_at	guanine nucleotide binding protein-like 3 (nucleolar)	GNL3	0.903
201739_at	serum/glucocorticoid regulated kinase	SGK	0.892
208152_s_at	DEAD (Asp-Glu-Ala-Asp) box polypeptide 21	DDX21	0.892
201466_s_at	v-jun sarcoma virus 17 oncogene homolog (avian)	JUN	0.887
205205_at	v-rel reticuloendotheliosis viral oncogene homolog B	RELB	0.887
206036_s_at	v-rel reticuloendotheliosis viral oncogene homolog	REL	0.886
218360_at	RAB22A, member RAS oncogene family	RAB22A	0.885
202241_at	tribbles homolog 1 (Drosophila)	TRIB1	0.884
212130_x_at	putative translation initiation factor	SUI1	0.879
214329_x_at	Tumour necrosis factor (ligand) superfamily, member 10	TNFSF10	0.878
64488_at	CDNA FLJ38849 fis, clone MESAN2008936	—	0.876
202871_at	TNF receptor-associated factor 4	TRAF4	0.873
207121_s_at	mitogen-activated protein kinase 6	MAPK6	0.870
200045_at	ATP-binding cassette, sub-family F (GCN20), member 1	ABCF1	0.863
202021_x_at	putative translation initiation factor	SUI1	0.862
202206_at	ADP-ribosylation factor-like 7	ARL7	0.861
202672_s_at	activating transcription factor 3	ATF3	0.856
204285_s_at	phorbol-12-myristate-13-acetate-induced protein 1	PMAIP1	0.849
204512_at	human immunodeficiency virus type I enhancer binding protein 1	HIVEP1	0.841
213376_at	zinc finger and BTB domain containing 1	ZBTB1	0.840
209325_s_at	regulator of G-protein signalling 16	RGS16	0.832
209959_at	nuclear receptor subfamily 4, group A, member 3	NR4A3	0.832
221877_at	CDNA FLJ38849 fis, clone MESAN2008936	—	0.816
220147_s_at	family with sequence similarity 60, member A	FAM60A	0.816
202963_at	regulatory factor X, 5 (influences HLA class II expression)	RFX5	0.815
209239_at	nuclear factor of kappa light polypeptide gene enhancer in B-cells 1	NFKB1	0.815
202207_at	ADP-ribosylation factor-like 7	ARL7	0.807
216396_s_at	etoposide induced 2.4 mRNA	EI24	0.800
217957_at	likely ortholog of mouse gene trap locus 3	GTL3	0.799
221571_at	TNF receptor-associated factor 3	TRAF3	0.797
201329_s_at	v-ets erythroblastosis virus E26 oncogene homolog 2 (avian)	ETS2	0.794

Continued on Next Page...

Table D.4 – Continued

Affymetrix tag	Description	Gene Symbol	Correlation
217833_at	CDNA FLJ31626 fis, clone NT2RI2003317	—	0.793
213618_at	centaurin, delta 1	CENTD1	0.791
211899_s_at	TNF receptor-associated factor 4	TRAF4	0.787
201631_s_at	immediate early response 3	IER3	0.786
204224_s_at	GTP cyclohydrolase 1 (dopa-responsive dystonia)	GCH1	0.784
207978_s_at	nuclear receptor subfamily 4, group A, member 3	NR4A3	0.783
219015_s_at	glycosyltransferase 28 domain containing 1	GLT28D1	0.782
205763_s_at	DEAD (Asp-Glu-Ala-Asp) box polypeptide 18	DDX18	0.781
204206_at	MAX binding protein	MNT	0.781
212458_at	sprouty-related, EVH1 domain containing 2	SPRED2	0.776
218647_s_at	ischemia/reperfusion inducible protein	YRDC	0.775
218023_s_at	family with sequence similarity 53, member C	FAM53C	0.771
201331_s_at	signal transducer and activator of transcription 6, interleukin-4 induced	STAT6	0.769
204286_s_at	phorbol-12-myristate-13-acetate-induced protein 1	PMAIP1	0.768
216060_s_at	dishevelled associated activator of morphogenesis 1	DAAM1	0.764
201079_at	synaptogyrin 2	SYNGR2	0.757
205479_s_at	plasminogen activator, urokinase	PLAU	0.757
212514_x_at	DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, X-linked	DDX3X	0.752
220890_s_at	DEAD (Asp-Glu-Ala-Asp) box polypeptide 47	DDX47	0.747
208726_s_at	eukaryotic translation initiation factor 2, subunit 2 beta	EIF2S2	0.746
207438_s_at	RNA, U transporter 1	RNUT1	0.746
221230_s_at	AT rich interactive domain 4B (RBP1- like)	ARID4B	0.741
212577_at	structural maintenance of chromosomes flexible hinge domain containing 1	SMCHD1	0.740
208804_s_at	splicing factor, arginine/serine-rich 6	SFRS6	0.739
204198_s_at	runt-related transcription factor 3	RUNX3	0.734
202076_at	baculoviral IAP repeat-containing 2	BIRC2	0.733
200880_at	DnaJ (Hsp40) homolog, subfamily A, member 1	DNAJA1	0.731
37028_at	protein phosphatase 1, regulatory (inhibitor) subunit 15A	PPP1R15A	0.730
220710_at	chromosome 15 open reading frame 28	C15orf28	0.730
213146_at	jumonji domain containing 3	JMJD3	0.729
207196_s_at	TNFAIP3 interacting protein 1	TNIP1	0.724
218558_s_at	mitochondrial ribosomal protein L39	MRPL39	0.723
203045_at	ninjurin 1	NINJ1	0.714
213097_s_at	zuotin related factor 1	ZRF1	0.710
41387_r_at	jumonji domain containing 3	JMJD3	0.707
208980_s_at	ubiquitin C	UBC	0.702
202510_s_at	tumour necrosis factor, alpha-induced protein 2	TNFAIP2	0.701

Table D.5: Ranked list of targets for training set 2

Affymetrix tag	Description	Gene Symbol	Correlation	Verification Score
201834_at	protein kinase, AMP-activated, beta 1 non-catalytic subunit	PRKAB1	0.973	6.30
203409_at	damage-specific DNA binding protein 2, 48kDa	DDB2	0.962	10.7
218346_s_at	sestrin 1	SESN1	0.934	3.90
216252_x_at	Fas (TNF receptor superfamily, member 6)	FAS	0.912	4.54
212371_at	CGI-146 protein	PNAS-4	0.897	2.61
218403_at	p53-inducible cell-survival factor	P53CSV	0.892	7.75
219361_s_at	hypothetical protein FLJ12484	FLJ12484	0.884	5.43
213293_s_at	tripartite motif-containing 22	TRIM22	0.879	6.07
218627_at	hypothetical protein FLJ11259	FLJ11259	0.875	3.56
205780_at	BCL2-interacting killer	BIK	0.875	6.57
215719_x_at	Fas (TNF receptor superfamily, member 6)	FAS	0.861	8.11
208796_s_at	cyclin G1	CCNG1	0.861	5.18
202181_at	KIAA0247	KIAA0247	0.828	2.22
201236_s_at	BTG family, member 2	BTG2	0.827	2.50
203725_at	growth arrest and DNA-damage-inducible, alpha	GADD45A	0.822	11.0
202284_s_at	cyclin-dependent kinase inhibitor 1A (p21, Cip1)	CDKN1A	0.816	8.07
201835_s_at	protein kinase, AMP-activated, beta 1 non-catalytic subunit	PRKAB1	0.813	5.92
204674_at	lymphoid-restricted membrane protein	LRMP	0.810	3.40
209295_at	tumour necrosis factor receptor superfamily, member 10b	TNFRSF10B	0.809	6.52
200730_s_at	protein tyrosine phosphatase type IVA, member 1	PTP4A1	0.798	4.45
219628_at	p53 target zinc finger protein	WIG1	0.798	3.70
207426_s_at	tumour necrosis factor (ligand) superfamily, member 4	TNFSF4	0.783	5.26
219627_at	hypothetical protein FLJ12700	FLJ12700	0.783	0.801
207616_s_at	TRAF family member-associated NFKB activator	TANK	0.779	-0.617
201093_x_at	succinate dehydrogenase complex, subunit A, flavoprotein (Fp)	SDHA	0.772	0.716
211318_s_at	RAE1 RNA export 1 homolog (S. pombe)	RAE1	0.771	3.44
204780_s_at	Fas (TNF receptor superfamily, member 6)	FAS	0.762	7.78
213038_at	IBR domain containing 3	IBRDC3	0.759	2.79
218527_at	aprataxin	APTX	0.754	-2.32
205692_s_at	CD38 antigen (p45)	CD38	0.750	9.02
214771_x_at	myosin phosphatase-Rho interacting protein	M-RIP	0.749	-0.935
202695_s_at	serine/threonine kinase 17a (apoptosis-inducing)	STK17A	0.748	6.50
218634_at	pleckstrin homology-like domain, family A, member 3	PHLDA3	0.748	3.71
200921_s_at	B-cell translocation gene 1, anti-proliferative	BTG1	0.745	5.42
208642_s_at	X-ray repair complementing defective repair in Chinese hamster cells 5	XRCC5	0.728	0.651
203846_at	tripartite motif-containing 32	TRIM32	0.721	0.862
203578_s_at	solute carrier family 7, member 6	SLC7A6	0.716	1.05
212815_at	activating signal cointegrator 1 complex subunit 3	ASCC3	0.716	5.93
204566_at	protein phosphatase 1D magnesium-dependent, delta isoform	PPM1D	0.714	6.05
209375_at	xeroderma pigmentosum, complementation group C	XPC	0.714	5.804
218014_at	pericentrin 1	PCNT1	0.709	0.294
218007_s_at	ribosomal protein S27-like	RPS27L	0.705	9.365
218031_s_at	checkpoint suppressor 1	CHES1	0.704	1.15

Table D.6: Ranked list of targets for training set 3

Affymetrix tag	Description	Gene Symbol	Correlation
205822_s_at	3-hydroxy-3-methylglutaryl-Coenzyme A synthase 1 (soluble)	HMGCS1	0.967
201791_s_at	7-dehydrocholesterol reductase	DHCR7	0.960
201790_s_at	7-dehydrocholesterol reductase	DHCR7	0.952
221750_at	3-hydroxy-3-methylglutaryl-Coenzyme A synthase 1 (soluble)	HMGCS1	0.943
216396_s_at	etoposide induced 2.4 mRNA	EI24	0.801
201801_s_at	solute carrier family 29 (nucleoside transporters), member 1	SLC29A1	0.800
217850_at	guanine nucleotide binding protein-like 3 (nucleolar)	GNL3	0.792
213097_s_at	zuotin related factor 1	ZRF1	0.786
202074_s_at	optineurin	OPTN	0.784
205205_at	v-rel reticuloendotheliosis viral oncogene homolog B, nuclear factor of kappa light polypeptide gene enhancer in B-cells 3	RELB	0.780
201631_s_at	immediate early response 3	IER3	0.767
212130_x_at	putative translation initiation factor	SUI1	0.765
208289_s_at	etoposide induced 2.4 mRNA	EI24	0.762
218187_s_at	hypothetical protein FLJ20989	FLJ20989	0.760
219648_at	likely ortholog of mouse dilute suppressor	DSU	0.760
210983_s_at	MCM7 minichromosome maintenance deficient 7 (S. cerevisiae)	MCM7	0.758
205763_s_at	DEAD (Asp-Glu-Ala-Asp) box polypeptide 18	DDX18	0.758
218558_s_at	mitochondrial ribosomal protein L39	MRPL39	0.757
217716_s_at	Sec61 alpha 1 subunit (S. cerevisiae)	SEC61A1	0.754
201947_s_at	chaperonin containing TCP1, subunit 2 (beta)	CCT2	0.744
202107_s_at	MCM2 minichromosome maintenance deficient 2, mitotin (S. cerevisiae)	MCM2	0.735
213376_at	zinc finger and BTB domain containing 1	ZBTB1	0.726
216237_s_at	MCM5 minichromosome maintenance deficient 5, cell division cycle 46	MCM5	0.719
208905_at	cytochrome c, somatic	CYCS	0.716
207339_s_at	lymphotoxin beta (TNF superfamily, member 3)	LTB	0.715
207622_s_at	ATP-binding cassette, sub-family F (GCN20), member 2	ABCF2	0.713
220147_s_at	family with sequence similarity 60, member A	FAM60A	0.713

Table D.7: Ranked list of targets for training set 4

Affymetrix tag	Description	Gene Symbol	Correlation
203214_x_at	cell division cycle 2, G1 to S and G2 to M	CDC2	0.955
218805_at	GTPase, IMAP family member 5	GIMAP5	0.947
209408_at	kinesin family member 2C	KIF2C	0.941
218039_at	nucleolar and spindle associated protein 1	NUSAP1	0.928
210559_s_at	cell division cycle 2, G1 to S and G2 to M	CDC2	0.927
218248_at	FLJ22794 protein	FLJ22794	0.926
204649_at	trophinin associated protein (tastin)	TROAP	0.920
219148_at	PDZ binding kinase	PBK	0.905
213186_at	zinc finger DAZ interacting protein 3	DZIP3	0.850
201292_at	topoisomerase (DNA) II alpha 170kDa	TOP2A	0.834
201663_s_at	SMC4 structural maintenance of chromosomes 4-like 1 (yeast)	SMC4L1	0.830
212021_s_at	antigen identified by monoclonal antibody Ki-67	MKI67	0.823
204817_at	extra spindle poles like 1 (S. cerevisiae)	ESPL1	0.816
209172_s_at	centromere protein F, 350/400ka (mitosin)	CENPF	0.805
201291_s_at	topoisomerase (DNA) II alpha 170kDa	TOP2A	0.797
202951_at	serine/threonine kinase 38	STK38	0.797
218662_s_at	chromosome condensation protein G	HCAP-G	0.792
219510_at	polymerase (DNA directed), theta	POLQ	0.783
210074_at	cathepsin L2	CTSL2	0.783
205176_s_at	integrin beta 3 binding protein (beta3-endonexin)	ITGB3BP	0.779
211519_s_at	kinesin family member 2C	KIF2C	0.777
210416_s_at	CHK2 checkpoint homolog (S. pombe)	CHEK2	0.775
203817_at	—	—	0.772
204079_at	tyrosylprotein sulfotransferase 2	TPST2	0.764
201310_s_at	chromosome 5 open reading frame 13	C5orf13	0.753
208983_s_at	platelet/endothelial cell adhesion molecule (CD31 antigen)	PECAM1	0.750
203408_s_at	special AT-rich sequence binding protein 1	SATB1	0.750
209773_s_at	ribonucleotide reductase M2 polypeptide	RRM2	0.749
207761_s_at	DKFZP586A0522 protein	DKFZP586A0522	0.749
206271_at	toll-like receptor 3	TLR3	0.734
219650_at	FLJ20105 protein	FLJ20105	0.733
214349_at	—	—	0.730
215067_x_at	peroxiredoxin 2	PRDX2	0.727
204026_s_at	ZW10 interactor	ZWINT	0.720
204825_at	maternal embryonic leucine zipper kinase	MELK	0.715
64064_at	GTPase, IMAP family member 5	GIMAP5	0.708
202580_x_at	forkhead box M1	FOXM1	0.707
202394_s_at	ATP-binding cassette, sub-family F (GCN20), member 3	ABCF3	0.704
209693_at	astrotactin 2	ASTN2	0.703
201990_s_at	cAMP responsive element binding protein-like 2	CREBL2	0.700

Table D.8: Ranked list of targets for training set 5.

Affymetrix tag	Description	Gene Symbol	Correlation
204359_at	fibronectin leucine rich transmembrane protein 2	FLRT2	0.974
204972_at	2'-5'-oligoadenylate synthetase 2, 69/71kDa	OAS2	0.950
205359_at	A kinase (PRKA) anchor protein 6	AKAP6	0.940
205291_at	interleukin 2 receptor, beta	IL2RB	0.934
216025_x_at	—	—	0.933
206776_x_at	acrosomal vesicle protein 1	ACRV1	0.928
210288_at	killer cell lectin-like receptor subfamily G, member 1	KLRG1	0.927
216661_x_at	cytochrome P450, family 2, subfamily C, polypeptide 9	CYP2C9	0.923
213873_at	—	—	0.922
201647_s_at	scavenger receptor class B, member 2	SCARB2	0.921
203899_s_at	calcitonin gene-related peptide-receptor component protein	RCP9	0.917
215479_at	Sema domain, transmembrane domain (TM), and cytoplasmic domain, 6A	SEMA6A	0.916
212970_at	Full-length cDNA clone CS0DC015YK09 of Neuroblastoma Cot 25	—	0.913
216034_at	suppressor of hairy wing homolog 1 (Drosophila)	SUHW1	0.912
58367_s_at	zinc finger protein 419	ZNF419	0.911
215126_at	CDNA FLJ42949 fis, clone BRSTN2006583	—	0.908
203184_at	fibrillin 2	FBN2	0.908
211405_x_at	interferon, alpha 17	IFNA17	0.907
220032_at	hypothetical protein FLJ21986	FLJ21986	0.907
51774_s_at	hypothetical protein LOC222070	LOC222070	0.905
211123_at	solute carrier family 5 (sodium iodide symporter), member 5	SLC5A5	0.905
211516_at	interleukin 5 receptor, alpha	IL5RA	0.904
220195_at	methyl-CpG binding domain protein 5	MBD5	0.904
209242_at	paternally expressed 3	PEG3	0.903
220446_s_at	carbohydrate (N-acetylglucosamine 6-O) sulfotransferase 4	CHST4	0.902
201646_at	scavenger receptor class B, member 2	SCARB2	0.898
209839_at	dynamin 3	DNM3	0.898
219813_at	LATS, large tumour suppressor, homolog 1 (Drosophila)	LATS1	0.896
210367_s_at	prostaglandin E synthase	PTGES	0.896
215040_at	Hypothetical protein FLJ11712	FLJ11712	0.894
219205_at	serine racemase	SRR	0.893
35150_at	CD40 antigen (TNF receptor superfamily member 5)	CD40	0.892
220398_at	MGC4170 protein	MGC4170	0.891
205782_at	fibroblast growth factor 7 (keratinocyte growth factor)	FGF7	0.889
209863_s_at	tumour protein p73-like	TP73L	0.888
204365_s_at	chromosome 2 open reading frame 23	C2orf23	0.887
217033_x_at	neurotrophic tyrosine kinase, receptor, type 3	NTRK3	0.887
209569_x_at	DNA segment on chromosome 4 (unique) 234 expressed sequence	D4S234E	0.887
121_at	paired box gene 8	PAX8	0.886
219287_at	potassium large conductance Ca-activated channel, subfamily M, β member 4	KCNMB4	0.886
210988_s_at	prune homolog (Drosophila)	PRUNE	0.884
220138_at	heart and neural crest derivatives expressed 1	HAND1	0.884
215582_x_at	MCM3 minichromosome maintenance deficient 3 associated protein	MCM3AP	0.883
215178_x_at	N-acylsphingosine amidohydrolase (acid ceramidase)-like	ASAHL	0.883
214933_at	calcium channel, voltage-dependent, P/Q type, alpha 1A subunit	CACNA1A	0.883
208259_x_at	interferon, alpha 7	IFNA7	0.882
221402_at	olfactory receptor, family 1, subfamily F, member 1	OR1F1	0.881
206846_s_at	histone deacetylase 6	HDAC6	0.881
216409_at	Acyl-CoA synthetase long-chain family member 6	ACSL6	0.879
213196_at	zinc finger protein 629	ZNF629	0.879
210796_x_at	sialic acid binding Ig-like lectin 6	SIGLEC6	0.879
203074_at	annexin A8	ANXA8	0.878
222376_at	Transcribed locus, moderately similar to XP_512541.1	—	0.878
214421_x_at	cytochrome P450, family 2, subfamily C, polypeptide 9	CYP2C9	0.877
217020_at	retinoic acid receptor, beta	RARB	0.876
222328_x_at	Maternally expressed 3	MEG3	0.876
205991_s_at	paired related homeobox 1	PRRX1	0.876
210331_at	HECT, C2 and WW domain containing E3 ubiquitin protein ligase 1	HECW1	0.876
214923_at	ATPase, H+ transporting, lysosomal 34kDa, V1 subunit D	ATP6V1D	0.875
215754_at	scavenger receptor class B, member 2	SCARB2	0.874
200856_x_at	nuclear receptor co-repressor 1	NCOR1	0.874
220232_at	stearoyl-CoA desaturase 4	SCD4	0.873
209765_at	a disintegrin and metalloproteinase domain 19 (meltrin beta)	ADAM19	0.873
204752_x_at	poly (ADP-ribose) polymerase family, member 2	PARP2	0.873
203999_at	—	—	0.872
71933_at	wingless-type MMTV integration site family, member 6	WNT6	0.871
214569_at	interferon, alpha 5	IFNA5	0.870
44783_s_at	hairy/enhancer-of-split related with YRPW motif 1	HEY1	0.870
203139_at	death-associated protein kinase 1	DAPK1	0.870
209594_x_at	pregnancy specific beta-1-glycoprotein 9	PSG9	0.869
219542_at	NIMA related kinase 11	NEK11	0.868

Continued on Next Page...

Table D.8 – Continued

Affymetrix tag	Description	Gene Symbol	Correlation
44673_at	sialoadhesin	SN	0.868
206101_at	extracellular matrix protein 2	ECM2	0.867
209851_at	KIAA0853	KIAA0853	0.866
206663_at	Sp4 transcription factor	SP4	0.866
213307_at	SH3 and multiple ankyrin repeat domains 2	SHANK2	0.866
204323_x_at	neurofibromin 1	NF1	0.866
215775_at	Thrombospondin 1	THBS1	0.865
210422_x_at	solute carrier family 11, member 1	SLC11A1	0.865
210319_x_at	msh homeo box homolog 2 (Drosophila)	MSX2	0.865
217130_at	chromosome 9 open reading frame 33	C9orf33	0.864
205189_s_at	Fanconi anemia, complementation group C	FANCC	0.864
216005_at	Tenascin C (hexabrachion)	TNC	0.864
214033_at	ATP-binding cassette, sub-family C (CFTR/MRP), member 6	ABCC6	0.864
205600_x_at	homeo box B5	HOXB5	0.863
213300_at	KIAA0404 protein	KIAA0404	0.863
213378_s_at	DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide 11	DDX11	0.863
216229_x_at	HLA complex group 2 pseudogene 7	HCG2P7	0.863
220452_x_at	Similar to acetyl-Coenzyme A synthetase 3	CECR7	0.863
204109_s_at	nuclear transcription factor Y, alpha	NFYA	0.862
209108_at	tetraspanin 6	TSPAN6	0.862
219891_at	pyroglutamyl-peptidase I	PGPEP1	0.862
211546_x_at	synuclein, alpha (non A4 component of amyloid precursor)	SNCA	0.861
221680_s_at	ets variant gene 7 (TEL2 oncogene)	ETV7	0.861
41856_at	Unc-5 homolog B (C. elegans)	UNC5B	0.859
220853_at	glycosyltransferase-like domain containing 1	GTDC1	0.858
208683_at	calpain 2, (m/II) large subunit	CAPN2	0.858